The phylogenetic range of bacterial and viral pathogens of vertebrates

Liam P. Shaw[1,2,a], Alethea D. Wang[1,3,a], David Dylus[4,5,6], Magda Meier[1,7], Grega Pogacnik[1], Christophe Dessimoz[4,5,6,8,9], Francois Balloux[1]

[a]Co-first authors

Affiliations:

[1] UCL Genetics Institute, University College London, London WC1E 6BT, UK

[2] Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK

[3] Canadian University Dubai, Sheikh Zayed Road, Dubai, United Arab Emirates

[4] Department of Computational Biology, University of Lausanne, CH-1015 Lausanne, Switzerland

[5] Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland

[6] Swiss Institute of Bioinformatics, Génopode, CH-1015 Lausanne, Switzerland

[7] Genetics and Genomic Medicine, University College London Institute of Child Health, 30 Guilford Street, London, WC1N 1EH, UK

[8] Centre for Life's Origins and Evolution, Department of Genetics Evolution & Environment, University College London, London WC1E 6BT, UK

[9] Department of Computer Science, University College London, London WC1E 6BT, UK

Correspondence:

Liam P. Shaw (liam.philip.shaw@gmail.com) and Francois Balloux (f.balloux@ucl.ac.uk)

Keywords:

Host range; Host jumps; Emerging infectious diseases; Phylogenetics; Zoonotic diseases

Abstract

Many major human pathogens are multi-host pathogens, able to infect other vertebrate species. Describing the general patterns of host-pathogen associations across pathogen taxa is therefore important to understand risk factors for human disease emergence. However, there is a lack of comprehensive curated databases for this purpose, with most previous efforts focusing on viruses. Here, we report the largest manually compiled host-pathogen association database, covering 2,595 bacteria and viruses infecting 2,656 vertebrate hosts. We also build a tree for host species using nine mitochondrial genes, giving a quantitative measure of the phylogenetic similarity of hosts. We find that the majority of bacteria and viruses are specialists infecting only a single host species, with bacteria having a significantly higher proportion of specialists compared to viruses. Conversely, multi-host viruses have a more restricted host range than multi-host bacteria. We perform multiple analyses of factors associated with pathogen richness per host species and the pathogen traits associated with greater host range and zoonotic potential. We show that factors previously identified as important for zoonotic potential in viruses—such as phylogenetic range, research effort, and being vector-borne—are also predictive in bacteria. We find that the fraction of pathogens shared between two hosts decreases with the phylogenetic distance between them. Our results suggest that host phylogenetic similarity is the primary factor for host-switching in pathogens.

42 # Introduction

43 Pathogens vary considerably in their host ranges. Some can only infect a single host species,
44 whereas others are capable of infecting a multitude of different hosts distributed across
45 diverse taxonomic groups. Multi-host pathogens have been responsible for the majority of
46 recent emerging infectious diseases in both human (Jones et al., 2008; Karesh et al., 2012;
47 Taylor, Latham, & Woolhouse, 2001; Woolhouse & Gowtage-Sequeria, 2005) and animal
48 populations (Cleaveland, Laurenson, & Taylor, 2001; Daszak, Cunningham, & Hyatt, 2000).
49 A number of studies have concluded that pathogens having a broad host range spanning
50 several taxonomic host orders constitute a higher risk of disease emergence, compared to
51 pathogens with more restricted host ranges (Cleaveland et al., 2001; Howard & Fletcher,
52 2012; Kreuder Johnson et al., 2015; McIntyre et al., 2014; Olival et al., 2017; Parrish et al.,
53 2008; Woolhouse & Gowtage-Sequeria, 2005).

54 An important biological factor that is likely to limit pathogen host-switching is the degree of
55 phylogenetic relatedness between the original and new host species. For a pathogen, closely
56 related host species can be considered akin to similar environments, sharing conserved
57 immune mechanisms or cell receptors, which increases the likelihood of pathogen 'pre-
58 adaptation' to a novel host. Barriers to infection will depend on the physiological similarity
59 between original and potential host species (Poulin & Mouillot, 2005), factors that can depend
60 strongly on host phylogeny. Indeed, the idea that pathogens are more likely to switch between
61 closely related host species has been supported by studies in several host-pathogen systems
62 (Davies & Pedersen, 2008; Faria, Suchard, Rambaut, Streicker, & Lemey, 2013; Streicker et
63 al., 2010; Waxman, Weinert, & Welch, 2014). The likelihood of infection of a target host has
64 also been found to increase as a function of phylogenetic distance from the original host in a

number of experimental infection studies (Gilbert & Webb, 2007; Longdon, Hadfield, Webster, Obbard, & Jiggins, 2011; Perlman & Jaenike, 2003; Russell et al., 2009).

Nevertheless, there are also numerous cases of pathogens switching host over great phylogenetic distances, including within the host-pathogen systems mentioned above. For example, a number of generalist primate pathogens are also capable of infecting more distantly-related primates than expected (Cooper et al., 2012). Moreover, for zoonotic diseases, a significant fraction of pathogens have host ranges that encompass several mammalian orders, and even non-mammals (Woolhouse & Gowtage-Sequeria, 2005). Interestingly, host jumps over greater phylogenetic distances may lead to more severe disease and higher mortality (Farrell & Davies, 2019). One factor that could explain why transmission into more distantly related new hosts occurs at all is infection susceptibility; some host clades may simply be more generally susceptible to pathogens (e.g. if they lack broad resistance mechanisms). Pathogens would therefore be able to jump more frequently into new hosts in these clades regardless of their phylogenetic distance from the original host. In support of this, experimental cross-infections have demonstrated that sigma virus infection success varies between different *Drosophila* clades (Longdon et al., 2011), and a survey of viral pathogens and their mammalian hosts found that host order was a significant predictor of disease status (Levinson et al., 2013).

While an increasing number of studies have described broad patterns of host range for various pathogens (see Table 1), most report only crude estimates of the breadth of host range, and there have been few attempts to systematically gather quantitatively explicit data on pathogen host ranges. As noted by Bonneaud, Weinert, & Kuijper (2019), most work on pathogen emergence has focused on viruses since these are the cause of many high-profile outbreaks (e.g. *Ebolavirus*), but it is plausible that the processes underlying emergence may be different

in bacterial pathogens. We thus have little understanding of the overall variation in host range both within and amongst groups of pathogens. This has limited our ability to examine how pathogen host range correlates with the emergence of infectious diseases. Here, we address this gap in the literature by considering both bacterial and viral pathogens in the same dataset.

In calculating host range, rather than using a taxonomic proxy for host genetic similarity (e.g. broad animal taxonomic orders (Kreuder Johnson et al., 2015; McIntyre et al., 2014; Woolhouse & Gowtage-Sequeria, 2005) or mammalian host order (Han, Kramer, & Drake, 2016) — see Table 1) we use a quantitative phylogenetic distance measure. Such an approach has been used in a recent set of papers (Albery, Eskew, Ross, & Olival, 2019; Guth, Visher, Boots, & Brook, 2019; Olival et al., 2017) which all use an alignment of the mitochondrial gene cytochrome *b* to build a maximum likelihood tree, constrained to the order-level topology of the mammalian supertree (Fritz, Bininda□Emonds, & Purvis, 2009). As noted by Albery et al. (Albery et al., 2019), this mammalian supertree has limited resolution at the species tips. Here, we therefore extend this approach in two ways: (a) we use a concatenated alignment of nine mitocondrial genes rather than one, and (b) we also consider non-mammalian vertebrates (see Methods). We believe this represents the most extensive vertebrate host tree built to-date, and should give the most precise quantitative measure of the true genetic similarity between hosts.

To summarise, here we combine a systematic literature review for bacterial and viral host ranges with a mitochondrial multigene host phylogeny (Figure 1). We compiled a database of 2,595 bacteria and viruses which infect 2,656 vertebrate host species. Our quantiative analysis represents by far the most comprehensive picture of known host-pathogen associations.

111 # Materials and Methods

112 ## Pathogen species

113 We focused on bacteria and viruses, as taken together they are the pathogen groups
114 responsible for the majority of the burden of communicable disease in humans: the combined
115 contribution of HIV/AIDS (viral), tuberculosis (bacterial), diarrheal diseases (predominantly
116 bacterial and viral), lower respiratory diseases (predominantly bacterial and viral) and
117 neonatal diseases (predominantly bacterial) made up 76.9% of all global disability-adjusted-
118 life-years (DALYs) lost due to communicable disease in 2017 (Global Burden of Disease
119 Collaborative Network, 2017). Bacteria and viruses also represent the two most diverse
120 groups in terms of total number of unique pathogen species recognized (Wardeh, Risley,
121 McIntyre, Setzkorn, & Baylis, 2015; Woolhouse & Gaunt, 2007) across both human hosts and
122 vertebrate animals.

123 Bacteria and viruses infectious to humans and animals were systematically compiled by going
124 through the complete taxonomic lists of known species from their respective authoritative
125 organizations. Bacterial species were drawn from the LPSN 2016 release (Euzéby & Parte,
126 2016); and viral species were drawn from the International Committee on Taxonomy of
127 Viruses (ICTV) 2015 release (ICTV, 2015). As such, our database is exhaustive and inclusive
128 of all known and taxonomically recognized bacteria and virus pathogens as of December
129 2016. Bacteria and virus species found in abiotic environments or only in non-vertebrate hosts
130 were not included in our database.

131 ## Pathogen metadata

132 We collected further metadata for each pathogen species. Where available, we used the NCBI
133 Genome Report for a species (last downloaded: 12th March 2019) to include the mean genome

134 size, number of genes, and GC content. We also annotated each pathogen for the presence of
135 known invertebrate vectors (i.e. whether they can be 'vector-borne'). For bacteria, we
136 additionally included information on Gram stain, bacterial motility, spore formation, oxygen
137 requirement, and cellular proliferation. These traits were collated primarily from the GIDEON
138 Guide to Medically Important Bacteria (Berger & Berger, 2016), but where information was
139 missing we also searched the primary literature. For viruses, we also included Baltimore
140 classifications from the ICTV Master Species List (ICTV, 2015). For the analysis including
141 host traits for direct comparison with Olival et al. (2017), we used their dataset of the number
142 of disease-related publications for species (Olival et al. searched ISI Web of Knowledge and
143 PubMed using the scientific binomial AND topic keyword: disease* OR virus* OR
144 pathogen* OR parasit*).

145 **Pathogen-host interactions**

146 Our literature search was designed to be as exhaustive and systematic as possible (Figure 1a).
147 We used Google Scholar to conduct a literature search to verify if each bacterial or viral
148 species was associated with a human or vertebrate animal host. Search terms consisted of the
149 pathogen species name and the keywords: 'infection', 'disease', 'human', 'animal', 'zoo',
150 'vet', 'epidemic' or 'epizootic'. At least one primary paper documenting the robust interaction
151 (i.e. infection) of the bacteria or virus species with a host species needed to be found in our
152 search for the association to be included in our database. In addition, several reputable
153 secondary sources were used to further validate the identified pathogen-host interactions: the
154 GIDEON Guide to Medically Important Bacteria (Berger & Berger, 2016); the Global
155 Mammalian Parasite Database (Nunn & Altizer, 2005); and the Enhanced Infectious Diseases
156 Database (EID2) (Wardeh et al., 2015) (eid2.liverpool.ac.uk/).

157 We aimed to manually read all publications found with Google Scholar searches using our
158 keyword search terms. However, as some pathogen species are extremely well-studied and
159 manual review of all returned publications was not possible, we decided to read only the first
160 ten pages of search results ordered by 'relevance' (equivalent to a limit of 200 publications).
161 Obviously, species with >200 results tend to be either well-studied pathogens (e.g.
162 '*Mycobacterium tuberculosis*' + 'infection': 62,900 results in 2016) or species with prolific
163 host ranges (e.g. '*Chlamydia psittaci*' + 'infection': 19,700 results in 2016). For these species
164 we cannot claim to have captured all known hosts with our manual review; i.e. we may not
165 have documented every single host species the pathogen has been recorded as infecting.
166 However, we are confident that we managed to reasonably approximate the full taxonomic
167 breadth of host range, since the first ten pages of results for these well-studied pathogens
168 usually contained specialized review papers listing the vertebrate host species in which
169 infections had been documented.

170 The majority of bacterial and viral pathogens in our database are known to cause disease
171 symptoms in at least one of their host species. However, in order to be as comprehensive as
172 possible, we considered as a pathogen any species for which there was *any* evidence that it
173 can cause symptomatic adverse infections under natural transmission conditions, even if rare,
174 including: cases where the relationship with host species is commonly asymptomatic, cases
175 where the relationship is only symptomatic in neonatal or immunocompromised individuals,
176 or where only a single case of infection has been recorded to date. Cases of deliberate
177 experimental infection of host species were excluded from our database as we judged that
178 these did not constitute natural evidence of a host-pathogen association.

179 A minority of bacterial and viral species in our database have not, to date, been shown to
180 cause any infectious symptoms in the host species they naturally infect. However,

181 characterizing symptoms in wild animal populations is difficult. Furthermore, these pathogens
182 are often very closely related to pathogens which are definitely known to cause disease in the
183 same hosts. For example, *Corynebacterium sphenisci* was isolated in a single study from
184 apparently healthy wild penguins (Goyache et al., 2003) but is related to species which are
185 pathogens across vertebrate hosts e.g. *Corynebacterium pseudotuberculosis*, the causative
186 agent of lymphadenitis. Therefore, we included all species apart from bacteria and viruses
187 which we considered to be clearly non-pathogens i.e. well-studied commensal or mutualistic
188 examples such as *Lactobacilli* in the human microbiome (Walter, 2008). Important
189 invertebrate vector species were also documented in our database, but our main analysis was
190 restricted to vertebrate hosts.

191 **Host species**

192 The taxonomic status of each host species identified in the primary literature was brought up
193 to date by identifying the current taxonomically valid species name using the ITIS Catalog of
194 Life (Rosokov et al., 2016) and the NCBI Taxonomy Database
195 (www.ncbi.nlm.nih.gov/taxonomy). In some cases, hosts were not identified to the species
196 level, but were retained in our database if they were identified to the family/order level and
197 there were no other host species from the same family/order infected by the same pathogen
198 species. In other cases, hosts were identified to the sub-species level (e.g. *Sus scrofa*
199 *domesticus*) if these sub-groups were economically and/or sociologically relevant.

200 The full compiled database contained 13,671 associations (Figure 1a), including invertebrate
201 hosts (n=305) as well as vertebrates (n=2,913). However, we restricted our host-relatedness
202 analysis to vertebrates for which we could construct a mitochondrial gene phylogeny (Figure
203 1b).

204 **Definition of zoonosis**

205 We classified a pathogen as zoonotic if it infected both humans and additional vertebrate
206 animals, including those shared but not known to be naturally transmissible among different
207 host species, unlike the WHO's definition of 'any disease or infection that is naturally
208 transmissible from vertebrate animals to humans and vice-versa' (WHO, 2019). This
209 definition includes species that mostly infect their various hosts endogenously or via the
210 environment (i.e. opportunistic pathogens) such as species in the bacterial genus *Actinomyces*.
211 We did this based on the observation that many new infectious diseases occur through cross-
212 species transmissions and subsequent evolutionary adaptation. Pathogens could also evolve to
213 become transmissible between host species. In addition, we are interested in how the overall
214 host range and host relatedness of a pathogen effects its likelihood of emergence and its
215 association with other pathogen characteristics. We did not classify a bacterial or viral species
216 as zoonotic if it had only been recognized outside of human infection in invertebrate hosts.

217 **Host phylogeny**

218 To infer a phylogenetic tree for all 3,218 vertebrate and invertebrate species, we relied on
219 nine mitochondrial genes: *cox2*, *cytb*, *nd3*, *12s*, *16s*, *nd2*, *co3*, *coi*, and *nadh4*. Our strategy
220 was as follows. First, we collected mitochondrial genes for species that had mitochondrial
221 gene submissions present in the NCBI database. For species without a mitochondrial gene
222 submission but where a whole genome was present, we extracted the genes by blasting the
223 genes of a taxonomically closely related species and then extracting the gene from the
224 resulting alignment. If no mitochondrial gene or whole genome submissions were available,
225 we used the NCBI taxonomy to approximate the species using a closely related species (using
226 either available genes or sequences extracted from genomes). Using this strategy and some

227 manual filtering, we were able to obtain mitochondrial gene sequences for 3,069 species
228 (including invertebrates).

229 We merged these genes in their distinct orthologous groups (OGs) using OMA (Altenhoff et
230 al., 2018). We used the nine largest OGs that had our expected nine genes as a basis for
231 alignment to ensure that alignment was conducted on high quality related sequences. We
232 aligned sequences for each OG separately using mafft (v7) with the options '--localpair --
233 maxiterate 1000' (Katoh & Standley, 2013). We then used MaxAlign (v1.1) (Gouveia-
234 Oliveira, Sackett, & Pedersen, 2007) to get the best aligning sequences from all sequences. In
235 order to produce more consistent alignment when only partial gene submissions were
236 available, we used the '--add' parameter of mafft to append all the residual sequences that
237 were part of a corresponding OG. Then, we concatenated all OGs and inferred the
238 phylogenetic tree using IQ-TREE (v1.5.5) with the options '-bb 1000' and the HKY+R10
239 model as identified by ModelFinder part of the IQTREE run (Hoang, Chernomor, von
240 Haeseler, Minh, & Vinh, 2018; Nguyen, Schmidt, von Haeseler, & Minh, 2015).

241 We observed a strong similarity between cophenetic distances for the 551 mammals in our
242 multi-mitochondrial gene tree that are also included in the *cytb* tree produced by Olival et al.
243 (Olival et al., 2017) and used by subsequent studies (Albery et al., 2019; Guth et al., 2019),
244 but there were some discrepancies (Supplementary Figure 1). We did not investigate these
245 further, but suspect they may have stemmed from us not constraining our phylogeny to an
246 existing order-level topology; that is, our phylogeny represents host genetic distances inferred
247 solely from available mitochondrial gene sequences, agnostic to any other information.
248 However, this tree appeared to be globally highly consistent with NCBI taxonomic ordering,
249 with only a small minority of species disrupting monophyly of groups (n=93, 3.1%). The
250 apparently incorrect placement of these species could have several possible explanations,

including: mislabelling in the database, poor sequence quality, or problems with the tree inference. After pruning the tree to include only vertebrate species (*n*=2,656, Figure 1c), a reduced fraction of species disrupted monophyly of groups (*n*=40, 1.5%). The analyses presented in the main text include species which disrupted monophyly.

Phylogenetic host breadth

Following Olival et al. (Olival et al., 2017) we define the Phylogenetic Host Breadth (PHB) of a pathogen as a function (*F*) of the cophenetic matrix of pairwise distances d_{ij} between its *N* hosts. Specifically, we take the function over the upper triangle of this (symmetric) matrix:

$$\mathrm{PHB}_F = F\left(\sum_{i<j} d_{ij}\right)$$

We found that the mean PHB for a pathogen (i.e. the average inter-host distance) was correlated with the maximum PHB (i.e. the largest inter-host distance) (Supplementary Figure 2), so there was little practical difference in using either measure. For simplicity, PHB refers to $\mathrm{PHB}_{\mathrm{mean}}$ unless otherwise stated. In fitting generalized additive models to predict zoonotic potential, models included different PHB quantities (see below).

Statistical analyses

Code for all analyses is available in our code and data repository (https://github.com/liampshaw/Pathogen-host-range). Here, we give a brief overview of our statistical methods.

Descriptive analyses of pathogen traits and host range. To summarise the dataset and make the most use of the traits we manually compiled, we performed separate analyses of different pathogen traits and their associations with host range using a specialist/generalist

271 distinction. We used Chi-squared tests to assess whether viruses and bacteria differed in their
272 proportion of specialist pathogens or vector-borne pathogens. We used Wilcoxon rank sum
273 tests to compare the distributions of GC content and genome size for specialists and
274 generalists within viruses and bacteria. We used Chi-squared tests to compare the distribution
275 proportion of generalists within subsets of bacteria based on lifestyle factors: motility, cellular
276 proliferation, spore formation, and oxygen requirement. Reported p-values are not corrected
277 for multiple testing. These results are best viewed as exploratory; some of the conclusions
278 were not retained when looking at the partial effects of predictors in a best-fit GAM,
279 controlling for the effects of other variables.

280 **Generalized additive models (GAMs).** We fitted GAMs to rank predictors using the
281 approach of Olival et al. (2017), adapting their previously published code. Their approach
282 uses smoothed spline predictors and automated double penalty smoothing to eliminate
283 redundant terms from the full model (select=TRUE in 'gam' function). Where multiple
284 variables measured the same effect, alternate GAMs used only one of these variables (e.g.
285 different proxies for research effort, although in practice these metrics are highly correlated).
286 Categorical and binary variables are fitted as random effects for each level of the variable.

287 **Host traits associated with greater pathogen richness.** We used a dataset of host traits for
288 terrestrial wild mammal species compiled by Olival et al. (2017) to predict pathogen richness
289 (bacterial and viral) per species. These host traits included a phylogenetic eigenvector
290 regression (PVR) of body mass. As Olival et al. collected information for an analysis of only
291 viral pathogens, we found there was better overlap for viruses in our dataset (n=613 host wild
292 mammals) than for bacteria (n=274).

Zoonotic potential. For bacteria, GAMs could include terms for host range ($\mathrm{PHB}_{\mathrm{mean}}$, $\mathrm{PHB}_{\mathrm{median}}$, or $\mathrm{PHB}_{\mathrm{max}}$), research effort (NCBI PubMed, Nucleotide, or SRA results), motility, sporulation, being vector-borne, oxygen requirements, and Gram stain. We excluded cellular lifestyle (intra/extracellular) as a predictor due to low numbers, and excluded pathogens of unknown motility (n=50) or sporulation (n=17). For viruses, GAMs could include terms for host range, research effort, genome size (number of proteins and length), being vector-borne, and genome type (Baltimore classification). We excluded pathogens with unknown genome size (n=253). We observed structure in some partial effect residuals in the best-fit GAMs: research effort for bacteria (Figure 6b) and host range for viruses (Figure 6d). This structure was driven by pathogen taxonomy, with families (orders) for bacteria (viruses) having different zoonotic potential; e.g. the *Staphylococcaceae* contain a high proportion of generalists. Attempts to include taxonomy as a categorical predictor produced best fit models which excluded all lifestyle factors (not shown), although host range and research effort were still the strongest predictors.

Shared pathogen analysis. If we denote the set of pathogens seen at least once in a host taxon a as p_a (where the taxon could be a species, genus, family etc.), we define the fraction of shared pathogens between two taxa a and b as:

$$s_{a,b} = \frac{|p_a \cap p_b|}{|p_a \cup p_b|}$$

Note that this definition is symmetric in *a, b.* It can therefore be compared with the (mean) phylogenetic distance between taxa using a Mantel test to determine the correlation.

Another property we consider is the fraction of the pathogens seen in a host taxon which are also seen in a reference host species (e.g. humans). Taking the comparison of primates and humans as an example: in the database, the primates (excluding humans) are represented by

147 host species with 762 host-pathogen associations. The total number of pathogen species with at least one association with a primate species is p_{P}=222. Of these, 158 are also seen in humans (the total number of pathogen species with at least one association with humans is p_{H}=1,675). We can then define the fraction of pathogens of primates also seen in humans as

$$f = \frac{|p_P \cap p_H|}{|p_P|} = \frac{158}{222} \sim 0.71$$

We used an illustrative sigmoidal fit of the form $y \sim A(1 + e^{\frac{(x_0 - x)}{s}})^{-1}$ separately for bacteria and viruses to model the relationship between *f* and the mean phylogenetic distance of the order to humans.

322 # Results

323 ## A comprehensive database of pathogen associations for vertebrates

324 Our database includes 12,212 associations between 2,595 vertebrate pathogens (1,685
325 bacterial species across 127 families; 910 viral species across 35 families) infecting 2,656
326 host species across 90 host orders. Pathogens infecting *Homo sapiens* made up 1,675 of all
327 associations (13.7%), the largest single host species group. The viral and bacterial pathogens
328 with the most recorded host associations were Newcastle disease virus (NDV) (n=207, 1.7%)
329 and *Chlamydia psittaci* (n=133, 1.1%), respectively.

330 ## Specialist pathogens are the most common category

331 Approximately half of all pathogens infected only a single host species (n=1,473, 56.8%;
332 Table 2). For pathogens not infecting humans, specialists were less common than generalists.
333 Bacteria had a significantly higher proportion of specialists compared to viruses (64.5% vs.
334 42.5%; χ^2 test: χ^2[df=1, n=1,635]=5.77, p=0.016). Almost half of all bacteria were human
335 specialists (855 of 1,685, 50.7%). Despite the dominance of specialists, many generalist
336 pathogens had broad host ranges spanning more than one host order: around one in three
337 pathogens infected multiple host orders (n=508, 30.3% of bacteria, n=307, 33.7% of viruses;
338 Table 2).

339 Considering well-represented pathogen taxonomic families (>20 pathogens in association
340 database), the bacterial family with the highest proportion of generalists was
341 *Staphylococcaceae* (24 of 29, 83%; Supplementary Figure 3). For viruses, it was
342 *Bunyaviridae* (68 of 79, 86%; Supplementary Figure 4).

343 **Multi-host viruses have a more restricted host range than multi-host bacteria**

344 Although the majority of pathogens infect just one host, and the total proportion of bacteria
345 and viruses infecting multiple host orders was similar (30.1% vs. 33.7%), the distribution of
346 generalists was significantly different between bacteria and viruses. Multi-host viruses were
347 more likely than bacteria to only infect a single host family (Table 2). A minority of
348 pathogens were vector-borne (n=272, 10.5%), and viruses were significantly more likely to be
349 vector-borne than bacteria (18.4% vs. 6.2%; χ^2 test: χ^2[df=1, n=2,589]=91.5, p<0.001). A
350 higher proportion of vector-borne viruses were generalists than those which were not vector-
351 borne (70.1% vs. 36.6%; χ^2 test: χ^2[df=1, n=907]=60.9, p<0.001). The same was true for
352 bacteria (49.5% vs. 21.4%; χ^2 test: χ^2[df=1, n=1,682]=42.3, p<0.001) (Supplementary Table
353 1).

354 This restricted host range of multi-host viruses was also apparent in the distribution of mean
355 phylogenetic host breadth (PHB) for multi-host pathogens (Figure 2). Bacteria generally had a
356 more positively skewed distribution of mean PHB compared to viruses (Figure 2; median
357 0.520 vs. 0.409, p<0.001 Wilcoxon rank sum test). Notably, these distributions were both
358 above the median maximum phylogenetic distance between hosts from the same order, which
359 was 0.323. The observation that bacteria had a more positively skewed distribution of mean
360 PHB was reproduced when subsampling to exclude human hosts for both domestic and non-
361 domestic hosts (see repository).

362 **Pathogen richness varies by host order**

363 Observed pathogen richness varied at the level of host order (Figure 3). Considering only host
364 species with an association with at least one bacterial species and one viral species, bacterial
365 and viral richness were strongly correlated (Spearman's ρ=0.57, p<0.001). The proportions of

366 these bacteria and viruses shared with humans were more weakly correlated (Spearman's
367 ρ=0.21, p<0.001).

368 We used a dataset of host traits for wild mammals previously compiled by Olival et al. (2017)
369 to find predictors of total bacterial and viral richness within a species using GAMs (Figure 4).
370 More than 60% of total deviance was explained by the best-fit GAMs (Table 3a and 3b) for
371 bacteria and viruses respectively). The number of disease-related citations for a host species
372 was the strongest predictor of the number of both bacterial and viral pathogens, accounting
373 for ~80% of relative deviance.

374 **Pathogen genome and host range**

375 We observed different distributions of pathogen genome GC content and genome size
376 depending on whether a pathogen was a specialist or a generalist (Supplementary Figure 5).
377 We had information on the number of proteins for n=657 viruses (72.2%, 5,815 associations).
378 While there was no significant correlation between the number of proteins in a virus genome
379 and mean PHB (Spearman's ρ=0.06, p=0.13), there was a significant positive correlation
380 between genome size and PHB (Spearman's ρ=0.23, p<0.001).

381 We had information on genome GC content and genome size for a similar proportion of
382 bacteria (n=1,135, 67.4%, 4,619 associations; n=550 species lacked data). However, there
383 was no significant correlation between bacterial genome size and PHB (Spearman's ρ=-0.05,
384 p=0.10), although specialists had a slightly larger genome size than generalists (means: 3.66
385 vs. 3.30 Mb; Wilcoxon rank sum test: W=140,610, p=0.007).

386 **Pathogen factors affecting host range of viruses**

387 **Genome composition.** Viruses with RNA genomes had a greater PHB than DNA viruses
388 (median: 0.238 vs. 0). Subsetting further, +ve-sense single-stranded RNA viruses (Baltimore
389 group V) had the greatest PHB (Supplementary Figure 6).

390 DNA viruses typically have much larger genomes than RNA viruses. We therefore fitted a
391 linear model for mean PHB using both DNA/RNA genome and genome size, with an
392 interaction term. Having an RNA genome and a larger genome were both significantly
393 associated with greater mean PHB in this linear model (t=6.11 and t=4.58 respectively,
394 p<0.001 for both, Supplementary Table 2), with a non-significant interaction between them
395 (p=0.36). In line with this, we found that the proportion of zoonotic viruses was higher for
396 RNA viruses (198 of 572, 34.6%) than DNA viruses (33 of 286, 11.5%), in agreement with
397 Olival et al. (2017) who found a similar proportion (41.6% vs. 14.1%).

398 **Pathogen factors affecting host range of bacteria**

399 We looked at the effect of bacterial lifestyle factors on the proportion of specialist and
400 generalist pathogens (Figure 7).

401 **Motility.** The majority of bacterial species in our database were non-motile (non-motile:
402 n=1,121, motile: n=514, not applicable: n=50 e.g. *Mycoplasmatales*). Motile bacteria were
403 more likely to infect multiple hosts compared to non-motile bacteria (27.2% vs. 21.7%; χ^2
404 test: χ^2[df=1, n=1,635]=5.77, p=0.016).

405 **Cellular proliferation.** Bacteria with an extracellular lifestyle (n=161) were not more likely
406 to infect multiple hosts compared to obligate (n=53) or facultative (n=93) intracellular
407 pathogens (χ^2 test: χ^2[df=2, n=307]=0.29, p=0.86). Combining motility and cellular
408 proliferation in a linear model suggested that neither was associated with greater mean PHB
409 (Supplementary Table 3).

410 **Spore formation.** Only a small number of bacterial pathogens were spore-forming (n=91),
411 and they did not have a significantly different number of generalists compared to non-spore-
412 forming bacteria.

413 **Oxygen requirement.** The oxygen requirements of bacteria were significantly associated
414 with the proportion of generalists in each group (χ^2 test: χ^2[df=2, n=1,573]=55.5, p<0.001).
415 Aerobic bacteria (n=648) were nearly twice as likely to infect multiple hosts compared to
416 anaerobic bacteria (n=343) (20.8% vs. 10.8%). However, facultatively anaerobic bacteria
417 (n=582) had an even higher proportion of species infecting multiple hosts (31.5%).

418 **Predicting zoonotic potential from pathogen traits**

419 We fitted GAMs to predict whether or not a pathogen was zoonotic using pathogen traits and
420 inspected the partial effects for each predictor in the best-fit model (Figure 5). Best-fit GAMs
421 could explain ~30% of total deviance (Table 3). We found that research effort and host range
422 (excluding human hosts) were the two strongest predictors of zoonotic potential, together
423 accounting for >70% of relative deviance. For bacteria, being facultatively anaerobic or
424 microaerophilic were significantly associated with zoonotic potential (Figure 5c, Table 3c) ;
425 for viruses, those with an RNA genome had greater zoonotic potential (Figure 5f, Table 3d).
426 Vector-borne pathogens had greater zoonotic potential for both bacteria and viruses.

427 **Pathogen sharing between hosts decreases with phylogenetic distance**

428 The proportion of total pathogens shared between host orders decreased with phylogenetic
429 distance (Figure 6a). Comparing vertebrate host orders specifically to *Homo sapiens* showed
430 that the closer an order was to humans, the greater the fraction of pathogens that were shared
431 for both bacteria and viruses, with an approximately sigmoidal relationship (Figure 6b). The
432 decrease in the fraction of shared pathogens was steeper for viruses than bacteria.

433 # Discussion

434 In this work, we have compiled the largest human-curated database of bacterial and viral
435 pathogens of vertebrates across 90 host orders. To date, this represents the most detailed and
436 taxonomically diverse characterization of pathogen host range. Using this database, we have
437 been able to conduct a detailed quantitative analysis of the overall distribution of host range
438 (host plasticity) across two major pathogen classes (together bacteria and viruses comprise the
439 majority of infectious diseases). We also use this database to examine the proportion of
440 pathogens shared among host orders.

441 We found that pathogen sharing was strongly correlated with the phylogenetic relatedness of
442 vertebrate hosts. This finding corroborates and generalises the observation by Olival et al.
443 (2017) for viral pathogens of mammalian hosts, as well as other studies using smaller taxon-
444 specific datasets (13, 15, 21). This suggests that phylogeny is a useful general predictor for
445 determining the 'spillover risk' (i.e. the risk of cross-species pathogen transmission) of
446 different pathogens into novel host species for both bacteria and viruses. Given the difficulty
447 in predicting the susceptibility of cross-species spillovers (Parrish et al., 2008) and the
448 restriction of most previous work to viruses, this finding is an important step in our
449 understanding of the broad factors underlying and limiting pathogen host ranges.

450 The underlying mechanisms by which phylogeny affects spillover risk still need to be more
451 closely examined. Pathogens are likely to be adapted to particular host physiologies (e.g. host
452 cell receptors and binding sites), which are expected to be more similar between genetically
453 closer host species. One mechanism by which a pathogen may be able to establish a broader
454 host range is by exploiting more evolutionarily conserved domains of immune responses
455 rather than immune pathways with high host species specificity. Such an association has been

456 shown among viruses for which the cell receptor is known (Woolhouse, 2002). Interestingly,
457 we found that the decrease in the fraction of shared pathogens with increasing phylogenetic
458 distance was steeper for viruses than bacteria, which suggests that bacterial pathogens, on
459 average, have higher host plasticity than viruses (i.e. a greater ability to infect a more
460 taxonomically diverse host range). Future studies could examine whether host cell receptors
461 for bacterial pathogens are more phylogenetically conserved compared with host cell
462 receptors for viral pathogens.

463 When examining the overall distribution of host ranges, we found a substantial fraction of
464 both bacterial and viral pathogens that have broad host ranges, encompassing more than one
465 vertebrate host order. The evolutionary selection of pathogens that have broad host ranges has
466 been a key hypothesis underpinning the emergence of new zoonotic diseases (Cleaveland et
467 al., 2001; Woolhouse & Gowtage-Sequeria, 2005), and mean PHB has previously been shown
468 to be the strongest predictor of the zoonotic potential of viral pathogens (Olival et al., 2017).
469 High pathogen host plasticity has also been found to be associated with both an increased
470 likelihood of secondary human-to human transmissibility and broader geographic spread
471 (Kreuder Johnson et al., 2015), both of which are traits linked to higher pandemic potential.
472 Given these observations, it may be useful to more closely monitor those pathogens with the
473 highest mean PHBs that have not yet been identified as zoonoses.

474 Several traits were found to be significantly associated with bacterial and viral host ranges.
475 For viruses, RNA viruses and larger genome size were independently associated with a
476 broader host range. This is in line with RNA viruses appearing particularly prone to infecting
477 new hosts and causing emerging diseases, something which has been attributed to their high
478 mutation rate (Holmes, 2010). The positive association between viral genome size and host

479 range might be due to pathogens specialising on a narrower range of hosts requiring a smaller
480 number of genes to fulfil their replication cycle.

481 For bacteria, motile and aerobic pathogens had a wider host range, with the largest number of
482 hosts found for facultative anaerobes, perhaps suggesting a greater ability to survive both
483 inside and outside hosts. Conversely, we did not find a strong association between genome
484 size and host range in bacteria; in fact, specialists had slightly larger genomes on average
485 compared to generalists. Since genome reduction through loss of genes is a well-recognised
486 signature of higher virulence in bacteria (Weinert & Welch, 2017), this suggest that
487 pathogenicity may be largely uncorrelated to host range in bacteria. It would be interesting to
488 further explore these relationships for obligate and facultative pathogens in the future.

489 We found a surprising lack of association between the expected 'intimacy' of host-pathogen
490 relationships (as judged with pathogen lifestyle factors) and host range. We identified more
491 single-host bacteria than viruses, which was the opposite of what we predicted going into this
492 study. One possibility is that bacterial pathogens may be more dependent on the host
493 microbiome i.e. their ability to infect other host species may be more contingent on the
494 existing microbial community, compared to viruses. However, we recognize that literature
495 bias could contribute to this conclusion, particularly for RNA viruses which are more difficult
496 to identify and diagnose than other infective agents. We also found that intra-cellular and
497 extra-cellular bacteria had roughly the same number of hosts despite our expectation for intra-
498 cellular bacteria to have a more narrow host range due to their higher expected intimacy with
499 their host. However, it should be noted that information about cellular proliferation was only
500 available for 18% (307 of 1,685) of all bacteria in the database, and this is a trait which can be
501 difficult to unambiguously characterize (Silva, 2012).

502 Previous studies of viral pathogens have shown that those that are vector-borne tend to have
503 greater host ranges — whether through higher host plasticity (Kreuder Johnson et al., 2015)
504 or higher mean PHB (Olival et al., 2017). We replicate this observation for both viruses and
505 bacteria, suggesting a strong and consistent effect of being vector-borne for a pathogen. We
506 also found that greater host range was associated with greater zoonotic potential for viral and
507 bacterial pathogens, complementing previous work restricted to viruses (Olival et al., 2017).

508 We controlled for research effort (number of publications, or number of SRA records) and
509 found that it was a strong predictor of both greater pathogen richness within a host species,
510 and the zoonotic potential of a pathogen. However, disentangling these factors is difficult.
511 There could be increased research efforts to study known zoonoses to identify them in
512 animals in order to establish possible 'reservoirs', giving a biased picture. However, this
513 could also partly be a consequence of the global distribution of humans and their propensity
514 to transmit pathogens to both wild and domestic species. There are multiple documented
515 cases of zoonotic pathogens having transmitted from humans to other animals, rather than the
516 other way around. Prominent examples include the ancestor of the agent of tuberculosis,
517 which humans likely transmitted to cows (Brosch et al., 2002; Mostowy et al. 2002), or the
518 multiple host jumps of *Staphylococcus aureus* from humans to cattle, poultry and rabbits
519 (Viana et al., 2015; Weinert et al., 2012). Such host jumps from humans to animals may
520 contribute to the pattern of zoonotic species having broader host ranges in particular for
521 pathogens at high prevalence in humans.

522 Our results have several further limitations. Our database was compiled from a
523 comprehensive synthesis of the available evidence in the literature about host-pathogen
524 associations. Our results are therefore necessarily biased by differences in research intensity
525 among both different pathogen and host species; or, viewed another way, they are a fair

526 reflection of the current state of knowledge in the literature. For example, specialist pathogens
527 of humans were the largest single group of bacterial species most likely because these have
528 been comparatively well-studied. Research effort targeted towards a pathogen primarily
529 reflects its medical and/or economic impact, and as such does not invalidate a classification
530 into specialist and generalist pathogens. We do not expect research effort to *a priori* bias
531 comparisons of host range between viruses and bacteria, or subsets therein defined by
532 particular traits such as GC content. Although we attempted to control for research effort in
533 our statistical analyses which did not depend on a binary 'specialist' vs. 'generalist'
534 distinction, the limitation of reflecting the current state of knowledge still applies to any
535 literature-based review and cannot be avoided.

536 We did not investigate in detail how geographical and ecological overlap between host
537 species affects pathogen sharing. We found that greater sympatry with other mammal species
538 (defined as ≥20% area of target species range) was a positive predictor of viral sharing but
539 was negatively correlated with bacterial sharing (Figure 4). Olival et al. (2017) concluded
540 ≥20% was the minimum threshold for viral sharing, but it may not be appropriate for bacteria.
541 Future work using generalized additive mixed models (GAMMs) will be necessary to
542 properly control for autocorrelation in pathogen sharing networks (Albery et al., 2019), but
543 we provide some thoughts here. Geographical overlap provides the necessary contact for host
544 switching to occur (Davies & Pedersen, 2008), and some authors have claimed that the rate
545 and intensity of contact may be "even more critical" than host relatedness in determining
546 switching (Parrish et al., 2008). In support of this, 'spillovers' over greater phylogenetic
547 distances are more common where vertebrates are kept in close proximity in zoos or wildlife
548 sanctuaries (Kreuder Johnson et al., 2015). Similarly, although multi-host parasites generally
549 infect hosts that are closely related rather than hosts with similar habitat niches (Clark &

550 Clegg, 2017), ecology and geography have been found to be key factors influencing patterns
551 of parasite sharing in primates (Cooper et al., 2012). Whereas, contact between two host
552 species clearly provides a necessary but not sufficient condition for direct host switching,
553 phylogenetic relatedness dictates the likely success of such a switch. Therefore, although the
554 relative importance of phylogeny and geography may depend on the specific context, our
555 observation of the strong dependence of pathogen sharing on phylogenetic distance across all
556 vertebrates emphasises that this is the primary underlying biological constraint.

557 We have substantially improved on previous efforts to assess pathogen host range by using
558 quantitative values based on alignment of multiple mitochondrial genes. However, our
559 definition of species for pathogens remains somewhat arbitrary as it follows existing
560 conventions. For example, in the *Mycobacterium tuberculosis* complex, the very closely
561 related lineages *M. tuberculosis* (*n*=26 host associations), *M. bovis* (*n*=78), and *M. africanum*
562 (*n*=3) are all treated as separate pathogens. Contrastingly, the extremely genetically diverse
563 complex grouped under *Salmonella enterica* subsp. *enterica* (*n*=44 associations) is treated as
564 a single pathogen. Exploring different measures of 'phylogenetic scale' to investigate these
565 questions could help: phylogenetic scale is precisely defined for nested clades, but for non-
566 nested and distantly-related clades it is less clear which measures to use (Graham, Storch, &
567 Machac, 2018). Developing a parallel phylogenetic framework for pathogens to complement
568 our host phylogenetic framework may be desirable, but challenging. An alignment of marker
569 genes is tractable for bacteria (e.g. by using ribosomal proteins (Hug et al., 2016)), but more
570 problematic for viruses, which have likely evolved on multiple independent occasions
571 (Krupovic & Koonin, 2017). Tracing the ancestors of viruses among modern cellular
572 organisms could represent another route to see if their host distribution reflects their
573 evolutionary past. Potentially, an alignment-free genetic distance method could be used

instead; as thousands more genomes become available for both pathogens and their hosts, such a method may be the optimal way to incorporate all known genomic information at a broad scale.

In conclusion, we have compiled the largest dataset of bacterial and viral pathogens of vertebrate host species to date. This is an important resource that has allowed us to explore different factors affecting the distribution of host range of vertebrate pathogens. While we are still some way off having a clear overall understanding of the factors affecting pathogen-host interactions, our results represent a substantial step in that direction. Maintaining such comprehensive datasets into the future is challenging but important, in order to ensure that all available knowledge is synthesized — rather than drawing conclusions only from well-studied pathogens, which likely represent the exceptions and not the norm.

Acknowledgements

We acknowledge financial support from the European Research Council (ERC) (grant ERC260801—BIG_IDEA to FB). DD and CD acknowledge the support of the Swiss National Science Foundation (grant 150654). We would also like to acknowledge the many public databases used in the construction of our own database and thank all creators. We thank Olival and colleagues for releasing comprehensive code for their reproducible analyses, allowing us to adapt it for our own purposes and easily compare results.

Data Accessibility Statement

Associated data and analysis code is available as a Github repository (https://github.com/liampshaw/Pathogen-host-range) and also archived on figshare (doi: 10.6084/m9.figshare.8262779). This includes: the pathogen-host association database; the

host phylogenetic tree; other datasets derived from them; and an Rmarkdown notebook which reproduces all analyses in this paper.

Image credits

Figure 1. Icons were downloaded from FlatIcon (flaticon.com) and were made by Maxim Basinski (tick/cross symbols) and Chanut (document symbol). Pathogen images (influenza, bacterium, adenovirus, HIV) were from the Bacteriology Virology image set from Servier Medical Art (smart.servier.com). Silhouettes were downloaded from PhyloPic (phylopic.org) and were made available under the Public Domain Dedication license (human and monkey) or a Creative Commons Attribution license: falcon (by Liftarn, CC BY-SA 3.0: creativecommons.org/licenses/by-sa/3.0), opossum (by S. Werning, CC BY 3.0: creativecommons.org/licenses/by/3.0).

Figure 3. Silhouettes were downloaded from PhyloPic. Images were either made available under the Public Domain Dedication license (Chiroptera, Cingulata, Peramelemorphia, Primates, Scandentia) or a Creative Commons Attribution license: Carnivora and Rodentia (by R. Groom, CC BY 3.0 and CC BY-SA 3.0 respectively); Didelphimorphia, Pilosa, Lagomorpha, and Diprotodonia (by S. Werning, CC BY 3.0); Euliptyphla (by C. Rebler, CC BY-SA 3.0); Perissodactyla (by J. A. Venter, H. H. T. Prins, D. A. Balfour & R. Slotow and vectorized by T. M. Keesey, CC BY 3.0); Probscidea (by T. M. Keesey, CC BY 3.0); Cetartiodactyla (by Zimices, CC BY 3.0).

Figure 6. Silhouettes were downloaded from PhyloPic. Images were either made available under the Public Domain Dedication license (Chiroptera) or the Creative Commons Attribution license: Primates (by B. Bocianowski, CC BY 3.0), Didelphimorphia (by S. Werning, CC BY 3.0), Falconiformes (by Liftarn, CC BY 3.0).

619 **Author contributions**

620 Using the CRediT taxonomy, the contributions of all authors to different stages of the project
621 were as follows: conceptualization (FB, ADW, LPS), data curation (ADW, DD, MM, GP,
622 LPS), formal analysis (ADW, LPS, DD), funding acquisition (FB, CD), investigation (ADW,
623 LPS, DD), methodology (ADW, LPS, DD, CD, FB), software (ADW, LPS, DD), supervision
624 (CD, FB), visualization (ADW, DD, LPS), writing – original draft (LPS, ADW, FB), writing
625 – review & editing (LPS, ADW, DD, CD, FB).

626 # References

627 Albery, G. F., Eskew, E. A., Ross, N., & Olival, K. J. (2019). Predicting the global
628 mammalian viral sharing network using phylogeography. *BioRxiv*, 732255. doi:
629 10.1101/732255

630 Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D.,
631 … Dessimoz, C. (2018). The OMA orthology database in 2018: retrieving
632 evolutionary relationships among all domains of life through richer web and
633 programmatic interfaces. *Nucleic Acids Research*, *46*(Database issue), D477–D485.
634 doi: 10.1093/nar/gkx1019

635 Berger, S., & Berger, D. S. (2016). *GIDEON Guide to Medically Important Bacteria: 2016*.
636 GIDEON Informatics, Incorporated.

637 Bonneaud, C., Weinert, L. A., & Kuijper, B. (2019). Understanding the emergence of
638 bacterial pathogens in novel hosts. *Philosophical Transactions of the Royal Society B:*
639 *Biological Sciences*, *374*(1782), 20180328. doi: 10.1098/rstb.2018.0328

640 Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., … Cole,
641 S. T. (2002). A new evolutionary scenario for the *Mycobacterium tuberculosis*
642 complex. *Proceedings of the National Academy of Sciences of the United States of*
643 *America*, *99*(6), 3684–3689. doi: 10.1073/pnas.052548299

644 Clark, N. J., & Clegg, S. M. (2017). Integrating phylogenetic and ecological distances reveals
645 new insights into parasite host specificity. *Molecular Ecology*, *26*(11), 3074–3086.
646 doi: 10.1111/mec.14101

647 Cleaveland, S., Laurenson, M. K., & Taylor, L. H. (2001). Diseases of humans and their
648 domestic mammals: pathogen characteristics, host range and the risk of emergence.
649 *Philosophical Transactions of the Royal Society of London. Series B, Biological*
650 *Sciences*, *356*(1411), 991–999. doi: 10.1098/rstb.2001.0889

651 Cooper, N., Griffin, R., Franz, M., Omotayo, M., Nunn, C. L., & Fryxell, J. (2012).
652 Phylogenetic host specificity and understanding parasite sharing in primates. *Ecology*
653 *Letters*, *15*(12), 1370–1377. doi: 10.1111/j.1461-0248.2012.01858.x

654 Daszak, P., Cunningham, A. A., & Hyatt, A. D. (2000). Emerging infectious diseases of
655 wildlife--threats to biodiversity and human health. *Science (New York, N.Y.)*,
656 *287*(5452), 443–449.

657 Davies, T. J., & Pedersen, A. B. (2008). Phylogeny and geography predict pathogen
658 community similarity in wild primates and humans. *Proceedings of the Royal Society*
659 *B: Biological Sciences*, *275*(1643), 1695–1701. doi: 10.1098/rspb.2008.0284

660 Euzéby, J., & Parte, A. C. (2016). List of prokaryotic names with standing in nomenclature
661 (LPSN). Retrieved from http://www.bacterio.net

662 Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G., & Lemey, P. (2013).
663 Simultaneously reconstructing viral cross-species transmission history and identifying
664 the underlying constraints. *Philosophical Transactions of the Royal Society of London.*
665 *Series B, Biological Sciences*, *368*(1614), 20120196. doi: 10.1098/rstb.2012.0196

666 Farrell, M. J., & Davies, T. J. (2019). Disease mortality in domesticated animals is predicted
667 by host evolutionary relationships. *Proceedings of the National Academy of Sciences*
668 *of the United States of America*, *116*(16), 7911–7915. doi: 10.1073/pnas.1817323116

Fritz, S. A., Bininda-Emonds, O. R. P., & Purvis, A. (2009). Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology Letters*, *12*(6), 538–549. doi: 10.1111/j.1461-0248.2009.01307.x

Gilbert, G. S., & Webb, C. O. (2007). Phylogenetic signal in plant pathogen-host range. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 4979–4983. doi: 10.1073/pnas.0607968104

Global Burden of Disease Collaborative Network. (2017). Global Burden of Disease Study 2017 (GBD 2017) Results. Retrieved January 8, 2020, from http://ghdx.healthdata.org/gbd-results-tool

Gouveia-Oliveira, R., Sackett, P. W., & Pedersen, A. G. (2007). MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics*, *8*, 312. doi: 10.1186/1471-2105-8-312

Goyache, J., Ballesteros, C., Vela, A. I., Collins, M. D., Briones, V., Hutson, R. A., … Fernández-Garayzábal, J. F. (2003). *Corynebacterium sphenisci* sp. nov., isolated from wild penguins. *International Journal of Systematic and Evolutionary Microbiology*, *53*(Pt 4), 1009–1012. doi: 10.1099/ijs.0.02502-0

Graham, C. H., Storch, D., & Machac, A. (2018). Phylogenetic scale in ecology and evolution. *Global Ecology and Biogeography*, *27*(2), 175–187. doi: 10.1111/geb.12686

Guth, S., Visher, E., Boots, M., & Brook, C. E. (2019). Host phylogenetic distance drives trends in virus virulence and transmissibility across the animal-human interface. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1782), 20190296. doi: 10.1098/rstb.2019.0296

692 Han, B. A., Kramer, A. M., & Drake, J. M. (2016). Global Patterns of Zoonotic Disease in
693 Mammals. *Trends in Parasitology*, *32*(7), 565–577. doi: 10.1016/j.pt.2016.04.007

694 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2:
695 improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*,
696 *35*(2), 518–522. doi: 10.1093/molbev/msx281

697 Holmes, E. C. (2010). The comparative genomics of viral emergence. *Proceedings of the*
698 *National Academy of Sciences*, *107*(suppl 1), 1742–1746. doi:
699 10.1073/pnas.0906193106

700 Howard, C. R., & Fletcher, N. F. (2012). Emerging virus diseases: can we ever expect the
701 unexpected? *Emerging Microbes & Infections*, *1*(12), e46. doi: 10.1038/emi.2012.47

702 Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., …
703 Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5),
704 16048. doi: 10.1038/nmicrobiol.2016.48

705 ICTV. (2015). Virus Taxonomy: 2015 Release. Retrieved from https://talk.ictvonline.org/

706 Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak,
707 P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990–993.
708 doi: 10.1038/nature06536

709 Karesh, W. B., Dobson, A., Lloyd-Smith, J. O., Lubroth, J., Dixon, M. A., Bennett, M., …
710 Heymann, D. L. (2012). Ecology of zoonoses: natural and unnatural histories. *Lancet*
711 *(London, England)*, *380*(9857), 1936–1945. doi: 10.1016/S0140-6736(12)61678-X

712 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version
713 7: improvements in performance and usability. *Molecular Biology and Evolution*,
714 *30*(4), 772–80. doi: 10.1093/molbev/mst010

715 Kreuder Johnson, C., Hitchens, P. L., Smiley Evans, T., Goldstein, T., Thomas, K., Clements,
716 A., … Mazet, J. K. (2015). Spillover and pandemic properties of zoonotic viruses with
717 high host plasticity. *Scientific Reports*, *5*, 14830. doi: 10.1038/srep14830

718 Krupovic, M., & Koonin, E. V. (2017). Multiple origins of viral capsid proteins from cellular
719 ancestors. *Proceedings of the National Academy of Sciences of the United States of*
720 *America*, *114*(12), E2401–E2410. doi: 10.1073/pnas.1621061114

721 Levinson, J., Bogich, T. L., Olival, K. J., Epstein, J. H., Johnson, C. K., Karesh, W., &
722 Daszak, P. (2013). Targeting surveillance for zoonotic virus discovery. *Emerging*
723 *Infectious Diseases*, *19*(5), 743–747. doi: 10.3201/eid1905.121042

724 Longdon, B., Hadfield, J. D., Webster, C. L., Obbard, D. J., & Jiggins, F. M. (2011). Host
725 phylogeny determines viral persistence and replication in novel hosts. *PLoS*
726 *Pathogens*, *7*(9), e1002260. doi: 10.1371/journal.ppat.1002260

727 McIntyre, K. M., Setzkorn, C., Wardeh, M., Hepworth, P. J., Radford, A. D., & Baylis, M.
728 (2014). Using open-access taxonomic and spatial information to create a
729 comprehensive database for the study of mammalian and avian livestock and pet
730 infections. *Preventive Veterinary Medicine*, *116*(3), 325–335. doi:
731 10.1016/j.prevetmed.2013.07.002

732 Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A., & Behr, M. A. (2002). Genomic
733 deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *The*
734 *Journal of Infectious Diseases*, *186*(1), 74–80. doi: 10.1086/341068

735 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and
736 effective stochastic algorithm for estimating maximum-likelihood phylogenies.
737 *Molecular Biology and Evolution*, *32*(1), 268–274. doi: 10.1093/molbev/msu300

738 Nunn, C., & Altizer, S. (2005). Global Mammal Parasite Database. Retrieved from
739 http://www.mammalparasites.org/

740 Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L., & Daszak, P.
741 (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature*,
742 *546*(7660), 646–650. doi: 10.1038/nature22975

743 Parrish, C. R., Holmes, E. C., Morens, D. M., Park, E.-C., Burke, D. S., Calisher, C. H., …
744 Daszak, P. (2008). Cross-species virus transmission and the emergence of new
745 epidemic diseases. *Microbiology and Molecular Biology Reviews: MMBR*, *72*(3),
746 457–470. doi: 10.1128/MMBR.00004-08

747 Perlman, S. J., & Jaenike, J. (2003). Infection success in novel hosts: an experimental and
748 phylogenetic study of Drosophila-parasitic nematodes. *Evolution; International*
749 *Journal of Organic Evolution*, *57*(3), 544–557.

750 Poulin, R., & Mouillot, D. (2005). Combining phylogenetic and ecological information into a
751 new index of host specificity. *The Journal of Parasitology*, *91*(3), 511–514. doi:
752 10.1645/GE-398R

753 Rosokov, Y., Abucay, L., Orrell, T., Nicolson, D., Bailly, N., Kirk, P. M., … Penev, L. (Eds.).
754 (2016). *Species 2000 & ITIS Catalogue of Life, 2016 Annual Checklist.* Retrieved
755 from http://www.catalogueoflife.org/annual-checklist/2016/

756 Russell, J. A., Goldman-Huertas, B., Moreau, C. S., Baldo, L., Stahlhut, J. K., Werren, J. H.,
757 & Pierce, N. E. (2009). Specialization and geographic isolation among *Wolbachia*

758 symbionts from ants and lycaenid butterflies. *Evolution; International Journal of*
759 *Organic Evolution*, *63*(3), 624–640. doi: 10.1111/j.1558-5646.2008.00579.x

760 Silva, M. T. (2012). Classical labeling of bacterial pathogens according to their lifestyle in the
761 host: inconsistencies and alternatives. *Frontiers in Microbiology*, *3*. doi:
762 10.3389/fmicb.2012.00071

763 Streicker, D. G., Turmelle, A. S., Vonhof, M. J., Kuzmin, I. V., McCracken, G. F., &
764 Rupprecht, C. E. (2010). Host phylogeny constrains cross-species emergence and
765 establishment of rabies virus in bats. *Science (New York, N.Y.)*, *329*(5992), 676–679.
766 doi: 10.1126/science.1188836

767 Taylor, L. H., Latham, S. M., & Woolhouse, M. E. (2001). Risk factors for human disease
768 emergence. *Philosophical Transactions of the Royal Society of London. Series B,*
769 *Biological Sciences*, *356*(1411), 983–989. doi: 10.1098/rstb.2001.0888

770 Viana, D., Comos, M., McAdam, P. R., Ward, M. J., Selva, L., Guinane, C. M., … Penadés,
771 J. R. (2015). A single natural nucleotide mutation alters bacterial pathogen host
772 tropism. *Nature Genetics*, *47*(4), 361–366. doi: 10.1038/ng.3219

773 Walter, J. (2008). Ecological role of *Lactobacilli* in the gastrointestinal tract: implications for
774 fundamental and biomedical research. *Applied and Environmental Microbiology*,
775 *74*(16), 4985–4996. doi: 10.1128/AEM.00753-08

776 Wardeh, M., Risley, C., McIntyre, M. K., Setzkorn, C., & Baylis, M. (2015). Database of
777 host-pathogen and related species interactions, and their global distribution. *Scientific*
778 *Data*, *2*, 150049. doi: 10.1038/sdata.2015.49

Waxman, D., Weinert, L. A., & Welch, J. J. (2014). Inferring host range dynamics from comparative data: the protozoan parasites of new world monkeys. *The American Naturalist*, *184*(1), 65–74. doi: 10.1086/676589

Weinert, L. A., & Welch, J. J. (2017). Why might bacterial pathogens have small genomes? *Trends in Ecology & Evolution*, *32*(12), 936–947. doi: 10.1016/j.tree.2017.09.006

Weinert, L. A., Welch, J. J., Suchard, M. A., Lemey, P., Allen, A., & Fitzgerald, J. R. (2012). Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biology Letters*, *8*(5), 829–832. doi: 10.1098/rsbl.2012.0290

WHO. (2019). WHO Health Topic page: Zoonoses. Retrieved from https://www.who.int/topics/zoonoses/en/

Woolhouse, M. E. J. (2002). Population biology of emerging and re-emerging pathogens. *Trends in Microbiology*, *10*(10), s3–s7. doi: 10.1016/S0966-842X(02)02428-9

Woolhouse, M. E. J., & Gaunt, E. (2007). Ecological origins of novel human pathogens. *Critical Reviews in Microbiology*, *33*(4), 231–242. doi: 10.1080/10408410701647560

Woolhouse, M. E. J., & Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*, *11*(12), 1842–1847. doi: 10.3201/eid1112.050997

798 # Tables and Figures

799 **Table 1. Notable previous studies of the host range of pathogens using a database**
800 **approach.**

Study	Database details	Methods	Pathogen types	Host range classification	Main conclusions
Taylor et al. (2001)	1,415 human pathogens	Literature review	Viruses and prions, bacteria and rickettsia, fungi, protozoa, helminths	Categorical: zoonotic or not	Found that 61% of human pathogens are zoonotic. First study identifying zoonotic pathogens as a major risk factor for human disease emergence.
Cleaveland et al. (2001)	1,922 human and domestic mammal (livestock and carnivore) pathogens	Literature review	Viruses and prions, bacteria and rickettsia, fungi, protozoa, helminths	Categorical: taxonomic grouping of mammal hosts (carnivores, ungulates, primates, bats, rodents and marine mammals) Simple quantitative: single or multiple host; human, domestic or wildlife hosts; hosts of one or more taxonomic orders	Found that 63% of pathogens infect multiple hosts, with multiple-host infections making up a higher proportion of domestic mammal pathogens than human pathogens. First study providing simple quantitative data on the host ranges of human and domestic mammal pathogens; and identifying the ability to infect multiple hosts (especially across taxonomic orders) as a risk factor for human and domestic mammal diseases emergence.
Woolhouse & Gowtage-Sequeria (2005)	1,407 human pathogens	Literature review	Viruses and prions, bacteria and rickettsia, fungi, protozoa, helminths	Categorical: zoonotic or not; type of nonhuman vertebrate host (broad categories: bats, carnivores, primates, rodents, ungulates, other mammals and nonmammals Simple quantititave: number of host types (0 – human only, 1, 2 or 3+ host types).	Zoonotic pathogens identified as major risk factor for human diseases emergence, with the fraction of emerging pathogen species increasing with the breadth of host range (number of host types).
McIntyre et al. (2014)	2,597 pathogen species across 47 mammalian and avian hosts (including humans and animals commonly used in Europe as food or kept as pets) 4,223 host-pathogen associations	Automated data mining of NCBI meta-data and semi-automated literature searches	Viruses and prions, bacteria and rickettsia, fungi, protozoa, helminths	Categorical: taxonomic grouping of mammal hosts (carnivores, ungulates, primates, bats, rodents and marine mammals) Simple quantitative: 1, 2, or 2+ host species	Pathogens having greater numbers of host species have increased odds of being a risk factor for disease emergence. Multiple-host infections make up a higher proportion of domestic mammal pathogens than human pathogens.

Kreuder Johnson et al. (2015)	162 zoonotic pathogens	Literature review	Viruses	Simple quantitative: viral host range (host plasticity) calculated as the total count of animal taxonomic orders and ecological groups recognized as hosts	Viruses with high host plasticity are more likely to amplify viral spillover by human-to-human transmission and have broader geographic spread.
Wardeh et al. (2015)	8,905 'cargo' (pathogen) species across 6,314 'carrier' (host) species (22,515 carrier/cargo interactions)	Automated data mining of NCBI databases (Taxonomy, Nucleotide, PubMed)	Viruses and prions, bacteria and rickettsia, fungi, protozoa, helminths	Not applicable	Verified individual species interactions from automated procedures against interactions identified by literature reviews.
Han et al. (2016)	Zoonotic pathogens of 27 orders of terrestrial mammals	Literature review	Viruses, bacteria, protozoa, helminths	Categorical: mammalian host order	Identified the proportion of zoonotic host species in each mammalian order (carnivores and rodents harbor the most zoonoses). Mammals carry more bacteria than any other pathogen type, followed by viruses.
Olival et al. (2017)	586 pathogens across 754 mammal species (2,805 host-pathogen associations)	Literature and database review	Viruses	Quantitative: phylogenetic host breadth calculated from two phylogenetic trees (mammal supertree; maximum likelihood cytB tree)	First study to show that the proportion of zoonotic viruses per species increases with host phylogenetic proximity to humans.
Albery et al. (2019)	250 pathogens across 4196 mammal species	Literature and database review: database originally collected by Olival et al. (2017)	Viruses	Quantitative: phylogenetic host breadth calculated from mammalian supertree	Host phylogenetic similarity and geographic overlap are strong predictors of viral sharing among mammal species.
Guth et al. (2019)	67 zoonotic pathogens across 278 mammalian hosts (420 host-pathogen associations)	Literature and database review	Viruses	Quantitative: distance from humans on a cytochrome b phylogenetic tree	Zoonotic viruses found in animal hosts with increased host phylogenetic distance from humans correlate positively with morbidity and mortality in humans, but negatively with human transmissibility.

This study	Initially: 2,678 pathogens across 3,218 host species (13,671 host-pathogen associations). Subsequently restricted to 2,595 pathogens across 2,656 vertebrate hosts (12,212 host-pathogen associations)	Literature and database review	Viruses, bacteria	Quantitative: phylogenetic host breadth calculated from host mitochondrial gene phylogeny	First study to show that the proportion of shared pathogens between vertebrate hosts (not just with respect to phylogenetic proximity to humans) decreases with increasing phylogenetic distance. First study to show that multi-host bacteria infect more diverse hosts than multi-host viruses. Zoonotic pathogens infect more non-human hosts than non-zoonotic. pathogens.

801

802 **Table 2. Summary of pathogen niche and specificity types**. A 'specialist' pathogen infects
803 only a single host, a 'generalist' more than one. Generalists are categorised according to
804 whether their hosts are within the same family (e.g. Bovidae), order (e.g. Artiodactyla), or
805 across orders. Percentages are of the total pathogen species of each type (bacteria or virus).

			% of total pathogens	
	Bacteria	**Virus**	**Bacteria**	**Virus**
Niche				
Animal-only	380	540	22.6	59.3
Human-only	855	133	50.7	14.6
Zoonotic	450	237	26.7	26.0
Specificity				
Specialist	1086	387	64.5	42.5
Animal-only	231	254	13.7	27.9
Human-only	855	133	50.7	14.6
Generalist	599	523	35.5	57.5
Within host family	49	132	2.9	14.5
Within host order	42	84	2.5	9.2
Across host orders	508	307	30.1	33.7

806

807 **Table 3. Summary of best-fit GAMs for predicting bacterial or viral richness per wild**
808 **mammal species, and the probability of a pathogen being zoonotic.**
809

Term	Value	Z statistic	Chi-sq statistic	P-value	Effective Degrees of Freedom	Total deviance explained	Relative deviance explained
(a) Bacterial richness per host species (wild mammals, n=215)						**74.70%**	
Intercept	1.28	23.24		**<0.001**			
Disease-related publications (ISI Web of Knowledge & PubMed, log)			584.96	**<0.001**	3.63		81.55%
Order: Rodentia			47.45	**<0.001**	1		7.88%
Mammal sympatry (>20% range overlap)			46.42	**<0.001**	1.15		2.65%
Geographic range (log)			10.89	0.117	4.13		1.88%
Order: Primates			10.36	**0.002**	0.97		1.80%
Order: Diprotodontia			7.03	**0.005**	0.89		1.69%
Order: Chiroptera			7.5	**0.003**	0.86		1.56%
Order: Didelphimorphia			2.81	0.058	0.78		0.67%
Order: Lagomorpha			1.19	0.174	0.61		0.32%
(b) Viral richness per host species (wild mammals, n=470)						**64.10%**	
Intercept	1.11	28.65		**<0.001**			
Disease-related publications (ISI Web of Knowledge & PubMed, log)			2884.3	**<0.001**	4.51		76.59%
Order: Primates			105.73	**<0.001**	0.99		7.52%
Geographic range (log)			119.01	**<0.001**	3.16		5.72%
Order: Cetartiodactyla			35.02	**<0.001**	0.95		2.93%
Mammal sympatry (>20% range overlap)			41.31	**<0.001**	2.96		2.61%
Order: Chiroptera			18.81	**<0.001**	0.97		2.05%
Order: Eulipotyphla			8.81	**0.002**	0.9		1.30%
Phylogenetically-corrected body mass (PVR)			5.73	0.193	3.2		0.80%
Order: Cingulata			1.1	0.144	0.51		0.18%
Order: Didelphimorphia			1.88	0.097	0.67		0.30%
(c) Zoonotic potential of bacteria (n=1,597 bacterial species)						**32.40%**	
Intercept	-1.46	-10.63		**<0.001**			
Host range (PHB median, log)			159.77	**<0.001**	4.82		47.81%
Research effort (PubMed results, log)			151.9	**<0.001**	2.59		42.04%
Vector-borne			50.49	**<0.001**	0.95		5.63%

Oxygen requirement: microaerophilic			17.75	**0.005**	0.92	2.04%
Oxygen requirement: facultatively anaerobic			12.04	**0.01**	0.9	1.79%
Motility			2.64	0.152	0.66	0.55%
Gram-positive			0.54	0.277	0.29	0.15%
(d) Zoonotic potential of viruses (n=657 viral species)						**30.10%**
Intercept	-1.76	-10.59		**<0.001**		
Host range (PHB mean, log)			68.13	**<0.001**	3.36	36.66%
Research effort (SRA results, log)			54.72	**<0.001**	3.65	33.33%
Vector-borne			41.05	**<0.001**	0.95	13.59%%
Genome: -ssRNA			28.51	**<0.001**	0.99	11.28%
Genome: +ssRNA			39.61	**0.003**	0.89	5.14%

810

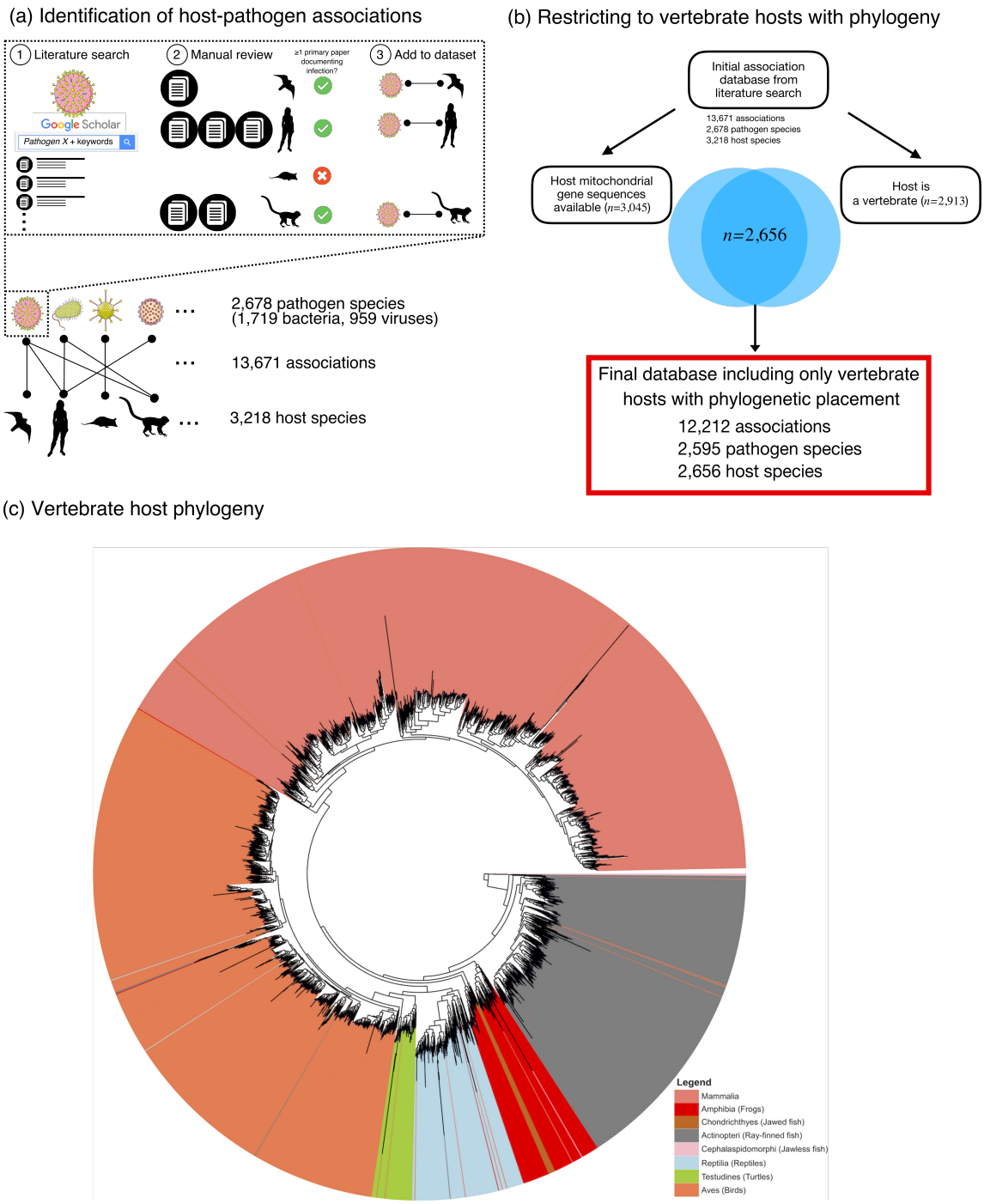
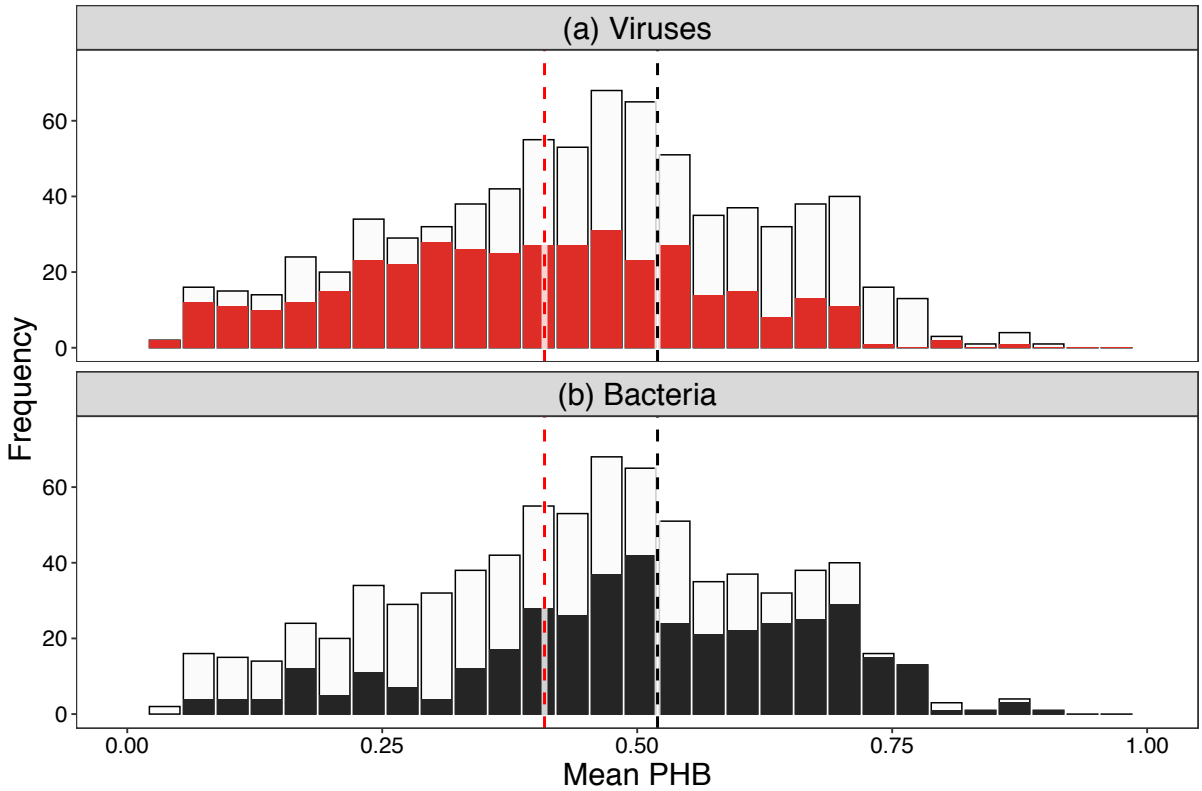


Figure 1. Overview of methodology for compiling the dataset. (a) Methodology of literature review (see Methods). (b) Subsetting the database to only associations involving vertebrate host species for which mitochondrial gene sequences could be identified. (c) Vertebrate host phylogeny using nine mitochondrial genes. (See 'Image credits' for credits and image licenses.)



817

818 **Figure 2. Mean phylogenetic host breadth for multi-host (top panel) bacteria and**
819 **(bottom panel) viruses.** Bacteria and viruses are shown in black and red respectively, with
820 the overall pathogen histogram (both types) shown in grey on both panels to help comparison.
821 On average, multi-host bacteria have a more diverse host range than viruses (black/red dashed
822 lines indicate median for bacteria/virus respectively). The majority of pathogens have a mean
823 PHB < 0.03 (*n*=1,816, 70.0%) and are excluded from the plot.

824

825

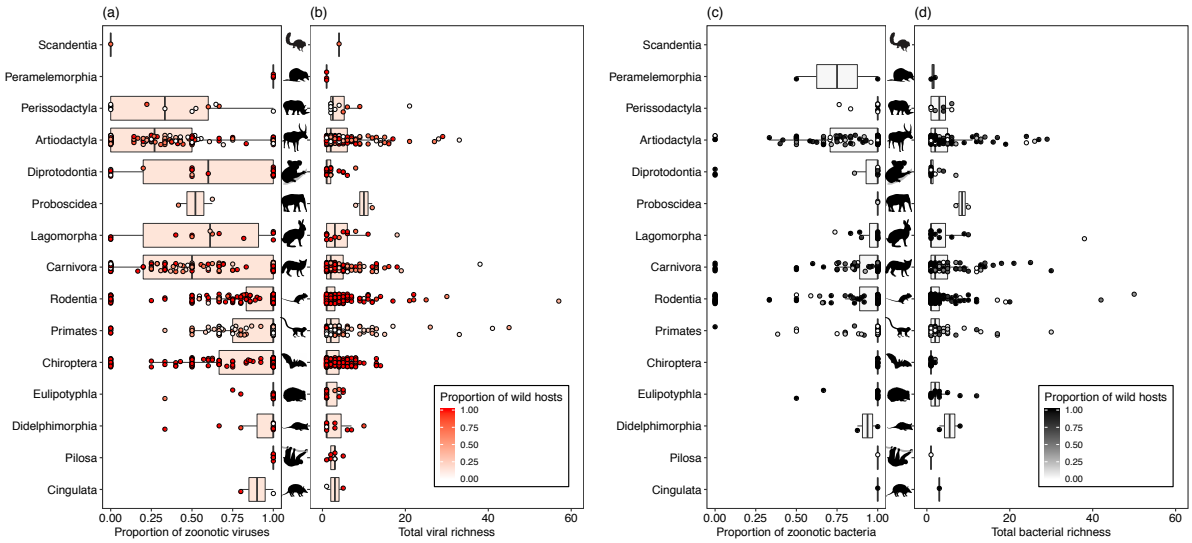


Figure 3. Observed pathogen richness in mammals. Box plots of (a,c) proportion of zoonotic viruses/bacteria and (b,d) total viral/bacterial richness per species, aggregated by order. Each point represents a host species, with the colour indicating the proportion of associations for a host species derived from observations of wild hosts (as opposed to in domestic or captive hosts of the same species). Lines represent median, boxes, interquartile range for each order. Data from 3,869 host–virus associations and 2,653 host-bacteria associations. The ordering of mammals is the same as in Figure 1 of Olival et al. (2017) to facilitate comparison. Richness plots have a cutoff at 60 to increase clarity of boxplots, so a small number of outlier species with greater pathogen richness are not shown (outliers for viral richness: one in Perissodactyla, four in Artiodactlya; outliers for bacterial richness: one in Perissodactyla, four in Artiodactyla, one in Carnivora). (See 'Image credits' for credits and image licenses.)

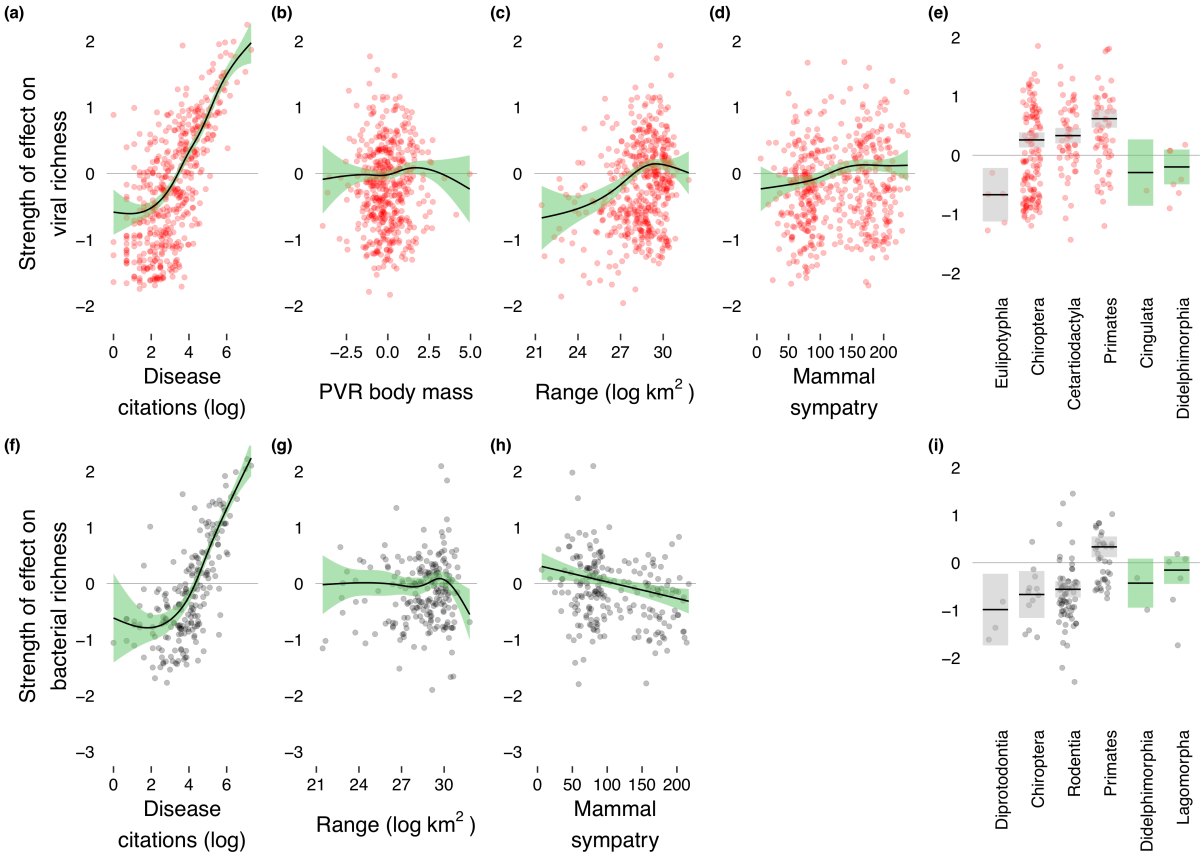


Figure 4. Host traits which predict total viral richness (top row) and bacterial richness (bottom row) per wild mammal species. Each plot shows the relative effect of the variable in the best-fit GAM accounting for the effects of other variables (see Table 3 for numerical values). Shaded circles represent partial residuals and shaded areas represent 95% confidence intervals around the mean partial effect. (a, f) Number of disease-related citations per host species, (b) phylogenetic eigenvector representation (PVR) of body mass i.e. corrected for phylogenetic signal, (c, g) geographic range area of host species, (d, h) number of sympatric mammal species overlapping with at least 20% area of target species range, and (e, i) mammalian order (non-significant terms retained in best model shown in grey). For bacteria, PVR body mass was not included in the best-fit GAM.

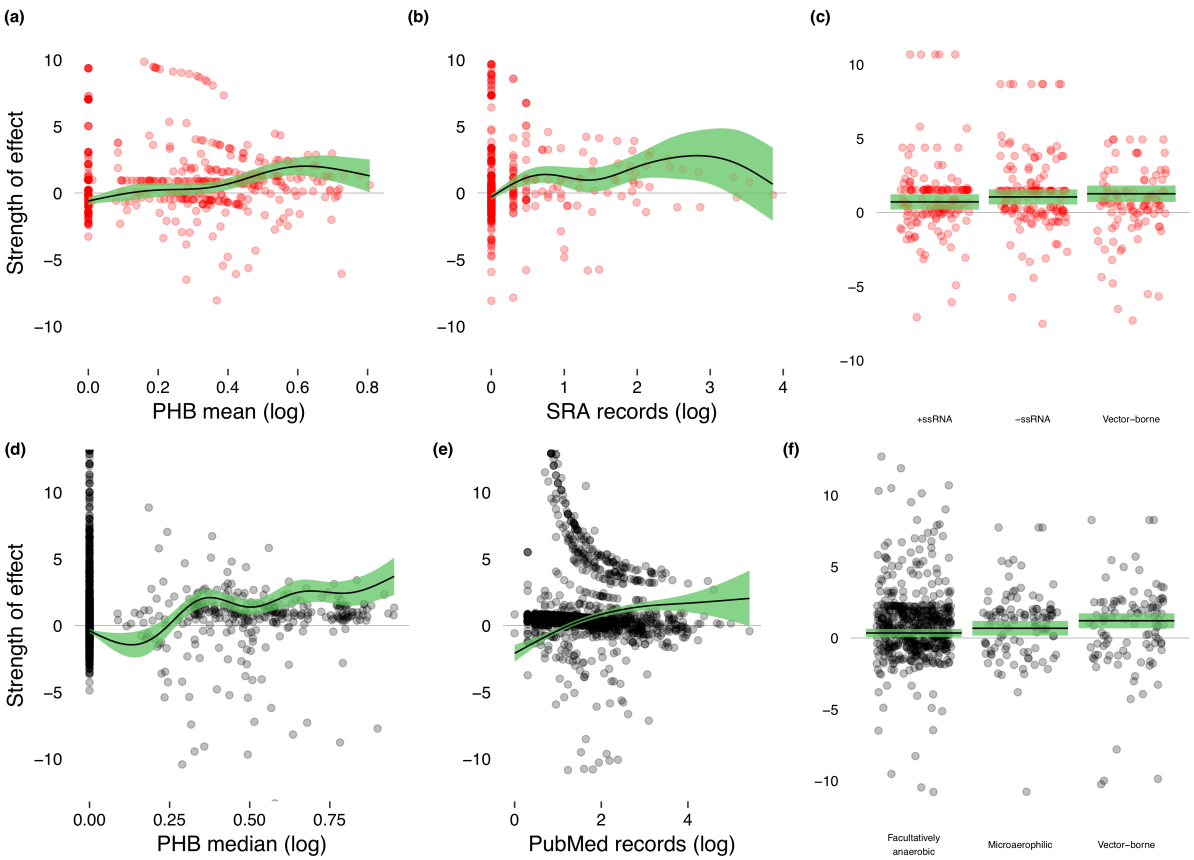


Figure 5. The partial effects of predictors in the best-fit GAM for predicting whether a bacterium (top row) or a virus (bottom row) is zoonotic. Each plot shows the relative effect of the variable in the best-fit GAM accounting for the effects of other variables (see Table 3 for numerical values). Shaded circles represent partial residuals and shaded areas represent 95% confidence intervals around the mean partial effect. Viruses: (a) Mean phylogenetic breadth of viral pathogen, (b) SRA records for viral pathogen, (c) Significant categorical predictors. +ssRNA and -ssRNA are mutually exclusive as they come from the 'genome type' variable. Bacteria: (d) Median phylogenetic breadth of bacterial pathogen, (e) PubMed records for bacterial pathogen, (f) Significant categorical predictors. Facultatively anaerobic and microaerophilic are mutually exclusive as they come from the 'oxygen' lifestyle variable. Gram stain and motility were included in the best-fit model but were not significant. Predictors are different for each best-fit GAM because the model term for e.g. phylogenetic breadth could be chosen from the mean, median, or maximum PHB for a pathogen.

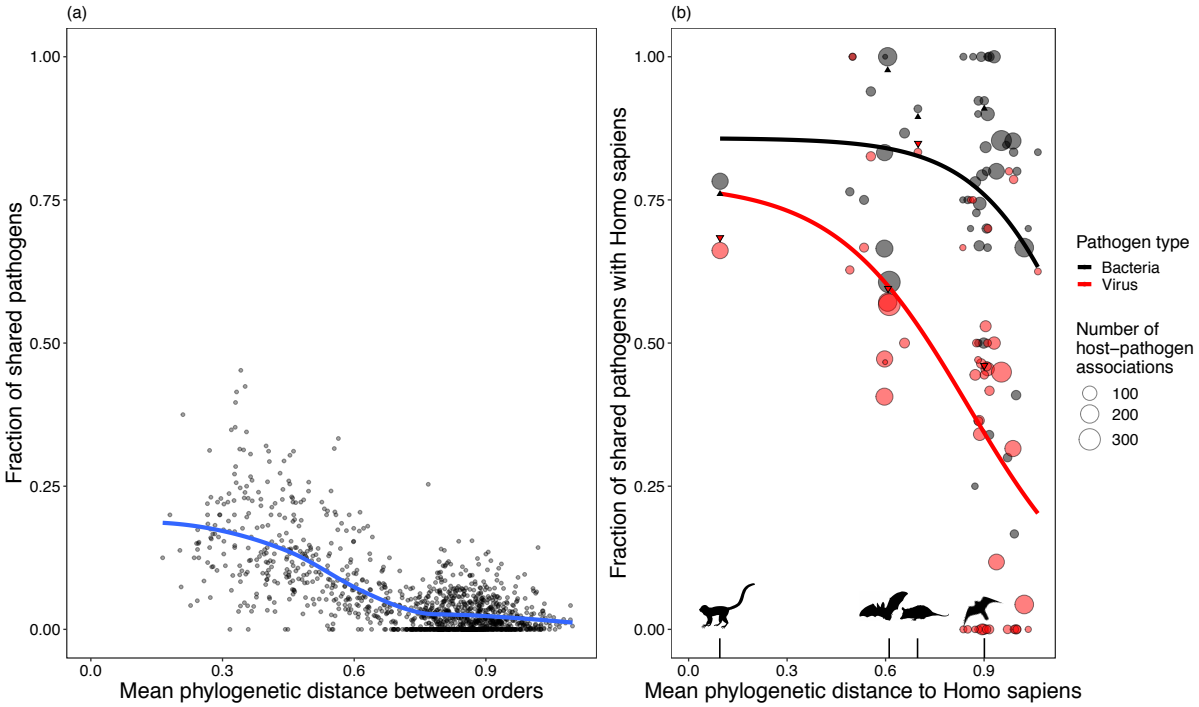


Figure 6. The fraction of shared pathogens between hosts decreases with inter-host phylogenetic distance. For the definitions of fraction of shared pathogens between two taxa, see Methods. (a) All pairwise comparisons between different host orders with at least 10 host-pathogen associations in the database. The blue line shows a smoothed average fit, produced with 'loess'. Only comparisons between host orders with at least 10 host-pathogen associations in the database are shown. (b) Fraction of shared pathogens between these host orders and *Homo sapiens* (as a fraction of total unique pathogens infecting at least one species in each order), showing data for bacteria (black) and viruses (red) together with a sigmoidal fit (thick line) for each pathogen type. Size of points indicates the number of unique host-pathogen associations for that order. Four illustrative orders are indicated with silhouettes (Primates, Chiroptera, Didelphimorphia, Falconiformes) with black (red) arrows indicating the relevant point for bacteria (viruses). (See 'Image credits' for credits and image licenses.)