bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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, United Kingdom

[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

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract

Pathogenic microorganisms can often infect more than one host. Furthermore, many major human pathogens are multi-host pathogens. Describing the general patterns of host-pathogen associations is therefore important to understand risk factors for human disease emergence. However, there is a lack of comprehensive curated databases for this purpose. Here, we report a manually compiled host-pathogen association database, covering 2,595 bacteria and viruses infecting 2,656 vertebrate hosts. We also built a multi-mitochondrial gene tree for host species, allowing us to show that the phylogenetic similarity of hosts is the dominant factor for greater pathogen sharing. We find that the majority of bacteria and viruses are specialists infecting only a single host species. Bacteria include a significantly higher proportion of specialists compared to viruses. Conversely, multi-host viruses have a more restricted host range than multi-host bacteria. Several traits are significantly associated with host range. For viruses, having an RNA genome and a larger viral genome size were independently associated with a broader host range. For bacteria, motile and aerobic pathogens had a wider host range, with the largest number of hosts found for facultative anaerobes. Unexpectedly, intracellular and extracellular bacteria had similar host ranges, despite *a priori* looser association of the latter with their hosts. We find that zoonotic pathogens typically have a larger phylogenetic range, and that the fraction of pathogens shared between two hosts decreases with the phylogenetic distance between them. This result suggests that host phylogenetic similarity is the primary factor for host-switching in pathogens.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Introduction

Pathogens vary considerably in their host ranges. Some can only infect a single host species, whereas others are capable of infecting a multitude of different hosts distributed across diverse taxonomic groups. Multi-host pathogens have been responsible for the majority of recent emerging infectious diseases in both human (1–4) and animal populations (5,6). Furthermore, a number of studies have concluded that pathogens having a broad host range spanning several taxonomic host orders constitute a higher risk of disease emergence, compared to pathogens with more restricted host ranges (2,6–11).

An important biological factor that is likely to limit pathogen host-switching is the degree of phylogenetic relatedness between the original and new host species. For a pathogen, closely related host species can be considered akin to similar environments, sharing conserved immune mechanisms or cell receptors, which increases the likelihood of pathogen 'pre-adaptation' to a novel host. Furthermore, barriers to infection will depend on the physiological similarity between original and potential host species (12), factors that can depend strongly on host phylogeny. Indeed, the idea that pathogens are more likely to switch between closely related host species has been supported by both correlative and experimental studies in several host-pathogen systems. Phylogenetic reconstructions of rabies virus transmissions have consistently shown decreasing rates of successful host shifts with increasing genetic distances between different bat hosts (13,14). Similarly, closely related primates have been shown to share more pathogens (15,16). The likelihood of infection of a target host has also been found to increase as a function of phylogenetic distance from the original host in a number of experimental infection studies. These include studies of sigma viruses (17), nematode parasites and their *Drosophila* hosts (18), fungal pathogens and their plant hosts (19), and *Wolbachia* and their insect hosts (20).

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 (21). Moreover, for zoonotic diseases, a significant fraction of pathogens have host ranges that encompass several mammalian orders, and even non-mammals (2). Interestingly, host jumps over greater phylogenetic distances may lead to more severe disease and higher mortality (22). 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 (17), and a survey of viral pathogens and their mammalian hosts found that host order was a significant predictor of disease status (23).

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. Similarly, there has been a lack of broad-scale comparative studies quantifying the degree of average host phylogenetic relatedness (using phylogenetic measures of host breadth) among 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.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

To address this gap in the literature, here we aim to quantify the phylogenetic range of vertebrate pathogens using a systematic literature review approach. We compiled a database of 2,595 bacteria and viruses which infect 2,656 vertebrate host species, representing by far the most comprehensive picture of recorded pathogen associations in the literature, allowing us to confidently draw general conclusions. Furthermore, rather than using a taxonomic proxy for host genetic similarity, we use a phylogenetic distance estimated from an alignment of all mitochondrial genes which gives a far more precise quantitative measure of the true genetic relatedness between hosts.

Results

A comprehensive database of pathogen associations for vertebrates

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

The majority of pathogens are specialists

The majority of pathogens infected only a single host species (56.7%, Table 2). For pathogens not infecting humans, specialists were less common than generalists. Bacteria had a significantly higher proportion of specialists compared to viruses (64.5% vs. 42.5%, p<0.001, χ^2 test). Almost half of all bacteria were human specialists (855 out of 1,685, 50.7%). Despite the dominance of specialists, many non-specialist pathogens had broad host ranges, with around one in three of all pathogens infecting multiple host orders.

Considering well-represented pathogen taxonomic families (>20 pathogens in association database), the bacterial family with the highest proportion of generalists was *Staphylococcaceae* (24 of 29, 83%; Supplementary Figure 1). For viruses, it was *Bunyaviridae* (68 of 79, 86%; Supplementary Figure 2).

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

Although the majority of pathogens infect just one host, and the total proportion of bacteria and viruses infecting multiple host orders was similar (30.1% vs. 33.7%), the distribution of generalists was significantly different between bacteria and viruses. Multi-host viruses were more likely than bacteria to only infect a single host family (Table 2). A minority of pathogens were vector-borne (n=272), and viruses were significantly more likely to be vector-borne than bacteria (18.4% vs. 6.2%, p<0.001, χ^2 test). A higher proportion of vector-borne viruses were generalists than those which were not vector-borne (66.4% vs. 36.6%, p<0.001, χ^2 test). The same was true for bacteria (49.5% vs. 21.4%, p<0.001, χ^2 test) (Supplementary Table 1).

This restricted host range of multi-host viruses was also apparent in the distribution of mean phylogenetic host breadth (PHB) for multi-host pathogens (Figure 2). Bacteria generally had a more positively skewed distribution of PHB compared to viruses (Figure 2; median 0.520 vs. 0.409, p<0.001 Wilcoxon rank sum test). Notably, these distributions were both above the median maximum phylogenetic distance between hosts from the same order, which was 0.323.

Pathogen genome and host range

We observed different distributions of pathogen genome GC content and genome size depending on whether a pathogen was a specialist or a generalist (Figure 3). We had

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

information on the number of proteins for 657 viruses (72.2%, 5,815 associations). While there was no significant correlation between the number of proteins in a virus genome and mean PHB (Spearman's ρ=0.06, p=0.13), there was a significant positive correlation between genome size and PHB (Spearman's ρ=0.23, p<0.001).

Conversely, there was no significant correlation between bacterial genome size and PHB (Spearman's ρ=-0.05, p=0.10), although specialists had a slightly larger genome size than generalists (mean: 3.66 vs. 3.30 Mb, p=0.007, Wilcoxon rank sum test).

Pathogen factors affecting host range of viruses

Genome composition. Viruses with RNA genomes had a significantly greater PHB than DNA viruses (median: 0.238 vs. 0, p<0.001, Wilcoxon rank sum test). Subsetting further, +ve-sense single-stranded RNA viruses (Baltimore group V) had the greatest PHB (Supplementary Figure 3).

DNA viruses typically have much larger genomes than RNA viruses. We therefore fitted a linear model for mean PHB using both DNA/RNA genome and genome size, with an interaction term. Having an RNA genome and a larger genome were both significantly associated with greater mean PHB (p<0.001 for both variables, Supplementary Table 2), with a non-significant interaction between them (p=0.36). In line with this, we found that the proportion of zoonotic viruses was higher for RNA (198 of 572, 34.6%) than DNA (33 of 286, 11.5%) viruses, in agreement with a previous virus-focused study (11) which found a similar proportion (41.6% vs. 14.1%).

Pathogen factors affecting host range of bacteria

We looked at the effect of bacterial lifestyle factors on the proportion of specialist and generalist pathogens (Figure 4).

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

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

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

Oxygen requirement. Aerobic bacteria (n=648) were nearly twice as likely to infect multiple hosts compared to anaerobic bacteria (n=343) (20.8% vs. 10.8%, p<0.001, χ^2 test). However, facultatively anaerobic bacteria (n=581) had an even higher proportion of species infecting multiple hosts (31.5%) which is significantly greater than for both aerobic and anaerobic bacteria (p<0.001, χ^2 test).

Pathogen sharing between hosts decreases with phylogenetic distance

The proportion of total pathogens shared between host orders decreased with phylogenetic distance (Figure 5a). Comparing vertebrate host orders specifically to *Homo sapiens* showed that there was also a significant correlation: the closer an order was to humans, the greater the fraction of pathogens that were shared for both bacteria and viruses (p<0.001; Figure 6b). The decrease in the fraction of shared pathogens was steeper for viruses than bacteria.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Zoonotic pathogens have a broader host range than non-zoonotic pathogens

Pathogens classified as zoonotic (infecting humans and at least one other vertebrate host, see Methods) had a significantly greater PHB even when only considering non-human hosts (Figure 6). This conclusion also held when considering the proportion of generalists ($p<0.001$ for both bacteria and viruses, χ^2 test).

Discussion

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

We found that pathogen sharing was strongly correlated with the phylogenetic relatedness of vertebrate hosts. This finding corroborates and generalises the observation by Olival et al. (11) for viral pathogens of mammalian hosts, as well as other studies using smaller taxon-specific datasets (13, 15, 21). This suggests that phylogeny is a useful general predictor for determining the 'spillover risk' (i.e. the risk of cross-species pathogen transmission) of different pathogens into novel host species. Given the previous lack of any apparent rules in predicting the susceptibility of cross-species spillovers (7), this finding is an important step in our understanding of the factors underlying and limiting pathogen host ranges.

The underlying mechanisms by which phylogeny affects spillover risk still need to be more closely examined. Pathogens are likely to be adapted to particular host physiologies (e.g. host cell receptors and binding sites), which are expected to be more similar between genetically closer host species. One mechanism by which a pathogen may be able to establish a broader host range is by exploiting more evolutionarily conserved domains of immune responses rather than immune pathways with high host species specificity. Such an association has been shown among viruses for which the cell receptor is known (24). Interestingly, we found that the decrease in the fraction of shared pathogens with increasing phylogenetic distance was steeper for viruses than bacteria, which suggests that bacterial pathogens, on average, have higher host plasticity than viruses (i.e. a greater ability to infect a more taxonomically diverse host range). Future studies could examine whether host cell receptors for bacterial pathogens are more phylogenetically conserved compared with host cell receptors for viral pathogens.

When examining the overall distribution of host ranges, we found a substantial fraction of both bacterial and viral pathogens that have broad host ranges, encompassing more than one vertebrate host order. The evolutionary selection of pathogens that have broad host ranges has been a key hypothesis underpinning the emergence of new zoonotic diseases (2,6), and mean PHB has previously been shown to be the strongest predictor of the zoonotic potential of viral pathogens (11). Furthermore, high pathogen host plasticity has also been found to be associated with both an increased likelihood of secondary human-to human transmissibility and broader geographic spread (10), both of which are traits linked to higher pandemic potential. Given these observations, it may be useful to more closely monitor those pathogens with the highest mean PHBs that have not yet been identified as zoonoses.

Several traits were found to be significantly associated with bacterial and viral host ranges. For viruses, RNA viruses and larger genome size were independently associated with a broader host range. This is in line with RNA viruses appearing particularly prone to infecting new hosts

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and causing emerging diseases, something which has been attributed to their high mutation rate (25). The positive association between viral genome size and host range might be due to pathogens specialising on a narrower range of hosts requiring a smaller number of genes to fulfil their replication cycle.

For bacteria, motile and aerobic pathogens had a wider host range, with the largest number of hosts found for facultative anaerobes, perhaps suggesting a greater ability to survive both inside and outside hosts. Conversely, we did not find a strong association between genome size and host range in bacteria; in fact, specialists had slightly larger genomes on average compared to generalists. Since genome reduction through loss of genes is a well-recognised signature of higher virulence in bacteria (26), this suggest that pathogenicity may be largely uncorrelated to host range in bacteria. It would be interesting to further explore these relationships for obligate and facultative pathogens in the future.

We found a surprising lack of association between the expected 'intimacy' of host-pathogen relationships (as judged with pathogen lifestyle factors) and host range. We identified more single-host bacteria than viruses, which was the opposite of what we would have predicted going into this study. One possibility is that bacterial pathogens may be more dependent on the host microbiome i.e. their ability to infect other host species may be more contingent on the existing microbial community, compared to viruses. However, we recognize that literature bias could contribute to this conclusion, particularly for RNA viruses which are more difficult to identify and diagnose than other infective agents. Furthermore, we also found that intra-cellular and extra-cellular bacteria had roughly the same number of hosts despite our expectation for intra-cellular bacteria to have a more narrow host range due to their higher expected intimacy with their host. However, it should be noted that information about cellular proliferation was only available for 18% (307 of 1,685) of all bacteria in the database, and this is a trait which can be difficult to unambiguously characterize (27).

Previous studies of viral pathogens have shown that those that are vector-borne tend to have greater host ranges — whether through higher host plasticity (10) or higher mean PHB (11) . We replicate this observation for both viruses and bacteria, suggesting a strong and consistent effect of being vector-borne. We also found that zoonotic multihost pathogens tend to have broader host ranges compared to non-zoonotic multihost pathogens. This result complements the previous finding that higher mean PHB was the strongest predictor of the zoonotic potential of viral pathogens of mammals (11). One potential caveat is that this finding could be driven by increased research efforts to study known zoonoses to identify them in animals in order to establish possible 'reservoirs', giving a biased picture. Conversely, this could partly be a consequence of the global distribution of humans and their propensity to transmit pathogens to both wild and domestic species. There are multiple documented cases of zoonotic pathogens having transmitted from humans to other animals, rather than the other way around. Prominent examples include the ancestor of the agent of tuberculosis, which humans likely transmitted to cows (28,29), or the multiple host jumps of *Staphylococcus aureus* from humans to cattle, poultry and rabbits (30,31). Such host jumps from humans to animals may contribute to the pattern of zoonotic species having broader host ranges in particular for pathogens at high prevalence in humans.

Our results have several further limitations. First, our database was compiled from a comprehensive synthesis of the available evidence in the literature about host-pathogen associations. Our results are therefore necessarily biased by differences in research intensity among both different pathogen and host species; or, viewed another way, they are a fair reflection of the current state of knowledge in the literature. For example, specialist pathogens of humans were the largest single group of bacterial species most likely because these have

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

been comparatively well-studied, not because any bacterial species chosen at random is likely to be a specialist human pathogen. These limitations apply to any literature-based review and will colour some of the results, such as the absolute number of pathogens per host species. We did not account for research effort in our study as we do not expect this to bias most 'relative patterns', such as comparison of host range between viruses and bacteria, or subsets defined by particular traits within these. Indeed, there is no reason to expect *a priori* that pathogens infecting humans would have particular traits.

We did not investigate how geographical and ecological overlap between host species affects pathogen sharing. Geographical overlap provides the necessary contact for host switching to occur (15), and some authors have claimed that the rate and intensity of contact may be "even more critical" than host relatedness in determining switching (7). In support of this, 'spillovers' over greater phylogenetic distances are more common where vertebrates are kept in close proximity in zoos or wildlife sanctuaries (10). Similarly, although multi-host parasites generally infect hosts that are closely related rather than hosts with similar habitat niches (32), ecology and geography have been found to be key factors influencing patterns of parasite sharing in primates (21). Contact between two host species clearly provides a necessary but not sufficient condition for direct host switching; phylogenetic relatedness dictates the likely success of such a switch. Therefore, although the relative importance of phylogeny and geography may depend on the specific context, our observation of the strong dependence of pathogen sharing on phylogenetic distance across all vertebrates emphasises that this is the general underlying biological constraint.

We have substantially improved on previous efforts to assess pathogen host range by using quantitative values based on alignment of the full complement of host mitochondrial genes. However, our definition of species for pathogens remains somewhat arbitrary as it follows existing conventions. For example, in the *Mycobacterium tuberculosis* complex, the very closely related lineages *M. tuberculosis* (n=26 host associations), *M. bovis* (n=78), and *M. africanum* (n=3) are all treated as separate pathogens. Contrastingly, the extremely genetically diverse complex grouped under *Salmonella enterica* subsp. *enterica* (n=44 associations) is treated as a single pathogen. Developing a parallel phylogenetic framework for pathogens to complement our host phylogenetic framework may be desirable, but challenging. An alignment of multiple marker genes is tractable for bacteria (e.g. by using ribosomal proteins (33)), but more problematic for viruses, which have likely evolved on multiple independent occasions (34). Tracing the ancestors of viruses among modern cellular organisms could represent another route to see if their host distribution reflects their 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.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Methods

Pathogen species

We focused on bacteria and viruses, as taken together they are the pathogen groups responsible for the majority of the burden of communicable disease in humans: the combined contribution of HIV/AIDS (viral), tuberculosis (bacterial), diarrheal diseases (predominantly bacterial and viral), lower respiratory diseases (predominantly bacterial and viral) and neonatal diseases (predominantly bacterial) made up 76.9% of all global disability-adjusted-life-years (DALYs) lost due to communicable disease in 2017 (35). Bacteria and viruses also represent the two most diverse groups in terms of total number of unique pathogen species recognized (36,37) across both human hosts and vertebrate animals.

Bacteria and viruses infectious to humans and animals were systematically compiled by going through the complete taxonomic lists of known species from their respective authoritative organizations. Bacterial species were drawn from the LPSN 2016 release (38); and viral species were drawn from the ICTV 2015 release (39). As such, our database is exhaustive and inclusive of all known and taxonomically recognized bacteria and virus pathogens as of December 2016.

Pathogen metadata

We collected further metadata for each pathogen species. Where available, we used the NCBI Genome Report for a species (last downloaded: 12th March 2019) to include the mean genome size, number of genes, and GC content. We also annotated each pathogen for the presence of known invertebrate vectors (i.e. whether they can be 'vector-borne'). For bacteria, we additionally included information on Gram stain, bacterial motility, spore formation, oxygen requirement, and cellular proliferation; for viruses, we included information on Baltimore classification (type of genome and method of classification).

Pathogen-host interactions

We used Google Scholar to conduct a literature search to verify if each bacterial or viral species was associated with a human or vertebrate animal host. Search terms consisted of the pathogen species name and the keywords: 'infection', 'disease', 'human', 'animal', 'zoo', 'vet', 'epidemic' or 'epizootic'. At least one primary paper documenting the robust interaction (i.e. infection) of the bacteria or virus species with a host species needed to be found in our search for the association to be included in our database. In addition, several reputable secondary sources were used to further validate the identified pathogen-host interactions: the GIDEON Guide to Medically Important Bacteria (40); the Global Mammalian Parasite Database (41); and the Enhanced Infectious Diseases Database (EID2) (42).

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

A minority of bacterial and viral species in our database have not, to date, been shown to cause any infectious symptoms in the host species they naturally infect e.g. the Torque teno virus (TTV). However, since they have yet to be classified as having commensal/mutualistic relationships with their host species, were included in our database. Bacteria and viruses that

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

are clearly commensal/mutualistic species were not included in our database. Important vector-hosts were also documented in our database, and were included in our host range analysis, but were not included in our analysis of host relatedness, which was restricted to vertebrate hosts.

Host species

Our literature search was designed to be as exhaustive and systematic as possible (Figure 1a). This was achieved by manually reading all Google Scholar hits obtained using our keyword search terms. However, as some pathogen species are extremely well-studied, we were unable to read through all associated publications for any pathogen that generated more than 10 pages of search results (equivalent to >200 publications). Obviously, these species tend to be either well-studied pathogens (e.g. *Mycobacterium tuberculosis*) or species with prolific host ranges (e.g. *Chlamydia psittaci*). For these species we cannot claim to have captured the entire numerical host range i.e. we may not have documented every single host species the pathogen has been recorded as infecting. However, we did attempt as far as possible to get a representative set covering the full taxonomic breadth of host range for each pathogen.

The taxonomic status of each host species identified in the primary literature was brought up to date by identifying the current taxonomically valid species name using the ITIS Catalog of Life (43) and the NCBI Taxonomy Database (44). In some cases, hosts were not identified to the species level, but were retained in our database if they were identified to the family/order level and there were no other host species from the same family/order infected by the same bacterial pathogen species. In other cases, hosts were identified to the sub-species level (e.g. *Sus scrofa domesticus*) if these sub-groups were economically and/or sociologically relevant.

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

Host phylogeny

To infer a tree for all 3,218 vertebrate and invertebrate species, we used a multi-mitochondrial gene approach using 9 genes: *cox2*, *cytb*, *nd3*, *12s*, *16s*, *nd2*, *co3*, *coi*, and *nadh4*. Our strategy was as follows. First, we collected mitochondrial genes for species that had mitochondrial gene submissions present in the NCBI database. For species without a mitochondrial gene submission but where a whole genome was present, we extracted the genes by blasting the genes of a taxonomically closely related species and then extracting the gene from the resulting alignment. If no mitochondrial gene or whole genome submissions were available, we used the NCBI taxonomy to approximate the species using a closely related species (using either available genes or sequences extracted from genomes). Using this strategy and some manual filtering, we were able to obtain mitochondrial gene sequences for 3,069 species (including invertebrates).

We merged these genes in their distinct orthologous groups (OGs) using OMA (45). We used the 10 largest OGs that had our expected 10 genes as basis for alignment to ensure that alignment was conducted on high quality related sequences. We aligned sequences for each OG separately using mafft (v7) with the options '--localpair --maxiterate 1000' (46). We then used MaxAlign (v1.1) (47) to get the best aligning sequences from all sequences. In order to produce more consistent alignment when only partial gene submissions were available, we used the '--add' parameter of mafft to append all the residual sequences that were part of a corresponding OG. Then, we concatenated all OGs and inferred the phylogenetic tree using IQ-TREE (v1.5.5) with the options '-bb 1000' and the HKY+R10 model as identified by ModelFinder part of the IQTREE run (48,49).

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The tree appeared to be globally highly consistent with NCBI taxonomic ordering, with only a small minority of species disrupting monophyly of groups (*n*=93, 3.1%). The 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 disrupted monophyly of groups (*n*=40, 1.5%). The analyses presented in the main text include these species; we found excluding them had no effect on our conclusions (see Supplementary Text 1).

Phylogenetic host breadth

Following Olival et al. (11) 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 was correlated with the maximum PHB (Supplementary Figure 4). We decided to choose to represent the phylogenetic range of a pathogen using its mean PHB i.e. PHB refers to $\mathrm{PHB}_{\mathrm{mean}}$ unless otherwise stated.

Definition of zoonosis

We classified a pathogen as zoonotic if it infected both human and vertebrate animals, including those shared but not known to be naturally transmissible among different host species, unlike the WHO's definition of "any disease or infection that is naturally transmissible from vertebrate animals to humans and vice-versa" (50). This definition includes species that mostly infect their various hosts endogenously or via the environment (i.e. opportunistic pathogens) such as species in the bacterial genus *Actinomyces*. We did this based on the observation that many new infectious diseases occur through cross-species transmissions and subsequent evolutionary adaptation. Pathogens could also evolve to become transmissible between host species. In addition, we are interested in how the overall host range and host relatedness of a pathogen effects its likelihood of emergence and its association with other pathogen characteristics. We did not classify a bacterial or viral species as zoonotic if it had only been recognized outside of human infection in invertebrate hosts.

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. In particular, we acknowledge use of the EID2 database funded primarily by the EU via the ERA ENV-HEALTH programme (ENHanCE project) and FP7 QWeCI and ICONZ projects.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Data availability

All associated data is available 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 (Supplementary Text 1).

References

1. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. [Online] 2001;356(1411): 983–989. Available from: doi:10.1098/rstb.2001.0888
2. Woolhouse MEJ, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*. [Online] 2005;11(12): 1842–1847. Available from: doi:10.3201/eid1112.050997
3. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. *Nature*. [Online] 2008;451(7181): 990–993. Available from: doi:10.1038/nature06536
4. Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, et al. Ecology of zoonoses: natural and unnatural histories. *Lancet (London, England)*. [Online] 2012;380(9857): 1936–1945. Available from: doi:10.1016/S0140-6736(12)61678-X
5. Daszak P, Cunningham AA, Hyatt AD. Emerging infectious diseases of wildlife--threats to biodiversity and human health. *Science (New York, N.Y.)*. 2000;287(5452): 443–449.
6. Cleaveland S, Laurenson MK, Taylor LH. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. [Online] 2001;356(1411): 991–999. Available from: doi:10.1098/rstb.2001.0889
7. Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, et al. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiology and molecular biology reviews: MMBR*. [Online] 2008;72(3): 457–470. Available from: doi:10.1128/MMBR.00004-08
8. Howard CR, Fletcher NF. Emerging virus diseases: can we ever expect the unexpected? *Emerging Microbes & Infections*. [Online] 2012;1(12): e46. Available from: doi:10.1038/emi.2012.47
9. McIntyre KM, Setzkorn C, Wardeh M, Hepworth PJ, Radford AD, Baylis M. Using open-access taxonomic and spatial information to create a comprehensive database for the study of mammalian and avian livestock and pet infections. *Preventive Veterinary Medicine*. [Online] 2014;116(3): 325–335. Available from: doi:10.1016/j.prevetmed.2013.07.002
10. Kreuder Johnson C, Hitchens PL, Smiley Evans T, Goldstein T, Thomas K, Clements A, et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific Reports*. [Online] 2015;5: 14830. Available from: doi:10.1038/srep14830
11. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. *Nature*. [Online] 2017;546(7660): 646–650. Available from: doi:10.1038/nature22975
12. Poulin R, Mouillot D. Combining phylogenetic and ecological information into a new index of host specificity. *The Journal of Parasitology*. [Online] 2005;91(3): 511–514. Available from: doi:10.1645/GE-398R

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

13. Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science (New York, N.Y.)*. [Online] 2010;329(5992): 676–679. Available from: doi:10.1126/science.1188836

14. Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. [Online] 2013;368(1614): 20120196. Available from: doi:10.1098/rstb.2012.0196

15. Davies T. Jonathan, Pedersen Amy B. Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proceedings of the Royal Society B: Biological Sciences*. [Online] 2008;275(1643): 1695–1701. Available from: doi:10.1098/rspb.2008.0284

16. Waxman D, Weinert LA, Welch JJ. Inferring host range dynamics from comparative data: the protozoan parasites of new world monkeys. *The American Naturalist*. [Online] 2014;184(1): 65–74. Available from: doi:10.1086/676589

17. Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. Host phylogeny determines viral persistence and replication in novel hosts. *PLoS pathogens*. [Online] 2011;7(9): e1002260. Available from: doi:10.1371/journal.ppat.1002260

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

19. Gilbert GS, Webb CO. Phylogenetic signal in plant pathogen-host range. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 2007;104(12): 4979–4983. Available from: doi:10.1073/pnas.0607968104

20. Russell JA, Goldman-Huertas B, Moreau CS, Baldo L, Stahlhut JK, Werren JH, et al. Specialization and geographic isolation among Wolbachia symbionts from ants and lycaenid butterflies. *Evolution; International Journal of Organic Evolution*. [Online] 2009;63(3): 624–640. Available from: doi:10.1111/j.1558-5646.2008.00579.x

21. Cooper N, Griffin R, Franz M, Omotayo M, Nunn CL, Fryxell J. Phylogenetic host specificity and understanding parasite sharing in primates. *Ecology Letters*. [Online] 2012;15(12): 1370–1377. Available from: doi:10.1111/j.1461-0248.2012.01858.x

22. Farrell MJ, Davies TJ. Disease mortality in domesticated animals is predicted by host evolutionary relationships. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 2019;116(16): 7911–7915. Available from: doi:10.1073/pnas.1817323116

23. Levinson J, Bogich TL, Olival KJ, Epstein JH, Johnson CK, Karesh W, et al. Targeting surveillance for zoonotic virus discovery. *Emerging Infectious Diseases*. [Online] 2013;19(5): 743–747. Available from: doi:10.3201/eid1905.121042

24. Woolhouse MEJ. Population biology of emerging and re-emerging pathogens. *Trends in Microbiology*. [Online] 2002;10(10): s3–s7. Available from: doi:10.1016/S0966-842X(02)02428-9

25. Holmes EC. The comparative genomics of viral emergence. *Proceedings of the National Academy of Sciences*. [Online] 2010;107(suppl 1): 1742–1746. Available from: doi:10.1073/pnas.0906193106

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

26. Weinert LA, Welch JJ. Why Might Bacterial Pathogens Have Small Genomes? *Trends in Ecology & Evolution*. [Online] 2017;32(12): 936–947. Available from: doi:10.1016/j.tree.2017.09.006
27. Silva MT. Classical Labeling of Bacterial Pathogens According to Their Lifestyle in the Host: Inconsistencies and Alternatives. *Frontiers in Microbiology*. [Online] 2012;3. Available from: doi:10.3389/fmicb.2012.00071 [Accessed: 28th May 2019]
28. Mostowy S, Cousins D, Brinkman J, Aranaz A, Behr MA. Genomic Deletions Suggest a Phylogeny for the Mycobacterium tuberculosis Complex. *The Journal of Infectious Diseases*. [Online] 2002;186(1): 74–80. Available from: doi:10.1086/341068
29. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 2002;99(6): 3684–3689. Available from: doi:10.1073/pnas.052548299
30. Weinert Lucy A., Welch John J., Suchard Marc A., Lemey Philippe, Rambaut Andrew, Fitzgerald J. Ross. Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an association with the spread of domestication. *Biology Letters*. [Online] 2012;8(5): 829–832. Available from: doi:10.1098/rsbl.2012.0290
31. Viana D, Comos M, McAdam PR, Ward MJ, Selva L, Guinane CM, et al. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nature Genetics*. [Online] 2015;47(4): 361–366. Available from: doi:10.1038/ng.3219
32. Clark NJ, Clegg SM. Integrating phylogenetic and ecological distances reveals new insights into parasite host specificity. *Molecular Ecology*. [Online] 2017;26(11): 3074–3086. Available from: doi:10.1111/mec.14101
33. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nature Microbiology*. [Online] 2016;1(5): 16048. Available from: doi:10.1038/nmicrobiol.2016.48
34. Krupovic M, Koonin EV. Multiple origins of viral capsid proteins from cellular ancestors. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 2017;114(12): E2401–E2410. Available from: doi:10.1073/pnas.1621061114
35. Global Burden of Disease Collaborative Network. *Global Burden of Disease Study 2017 (GBD 2017) Results.* [Online] Available from: http://ghdx.healthdata.org/gbd-results-tool
36. Woolhouse M, Gaunt E. Ecological origins of novel human pathogens. *Critical Reviews in Microbiology*. [Online] 2007;33(4): 231–242. Available from: doi:10.1080/10408410701647560
37. Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M. Database of host-pathogen and related species interactions, and their global distribution. *Scientific Data*. [Online] 2015;2: 150049. Available from: doi:10.1038/sdata.2015.49
38. Euzéby J, Parte AC. *List of prokaryotic names with standing in nomenclature (LPSN)*. [Online] Available from: http://www.bacterio.net
39. *International Committee on Taxonomy of Viruses (ICTV)*. [Online] Available from: https://talk.ictvonline.org/
40. GIDEON Informatics Inc, Berger DS. *GIDEON Guide to Medically Important Bacteria: 2016*. GIDEON Informatics, Incorporated; 2016. 1888 p.
41. Nunn C, Altizer S. *Global Mammal Parasite Database*. [Online] Available from: http://www.mammalparasites.org/

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

42. *Enhanced Infectious Diseases Database (EID2)*. [Online] Available from: https://eid2.liverpool.ac.uk/
43. Rosokov Y, Abucay L, Orrell T, Nicolson D, Bailly N, Kirk PM, et al., editors. *Species 2000 & ITIS Catalogue of Life, 2016 Annual Checklist.* [Online] Leiden, The Netherlands: Species 2000: Naturalis; Available from: http://www.catalogueoflife.org/annual-checklist/2016/
44. NCBI. *NCBI Taxonomy Database*. [Online] Available from: www.ncbi.nlm.nih.gov/taxonomy
45. Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*. [Online] 2018;46(Database issue): D477–D485. Available from: doi:10.1093/nar/gkx1019
46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*. [Online] 2013;30(4): 772–80. Available from: doi:10.1093/molbev/mst010
47. Gouveia-Oliveira R, Sackett PW, Pedersen AG. MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics*. [Online] 2007;8: 312. Available from: doi:10.1186/1471-2105-8-312
48. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*. [Online] 2015;32(1): 268–274. Available from: doi:10.1093/molbev/msu300
49. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*. [Online] 2018;35(2): 518–522. Available from: doi:10.1093/molbev/msx281
50. WHO. *WHO Health Topic page: Zoonoses*. [Online] Available from: https://www.who.int/topics/zoonoses/en/
51. Han BA, Kramer AM, Drake JM. Global Patterns of Zoonotic Disease in Mammals. *Trends in Parasitology*. [Online] 2016;32(7): 565–577. Available from: doi:10.1016/j.pt.2016.04.007

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Figures and Tables

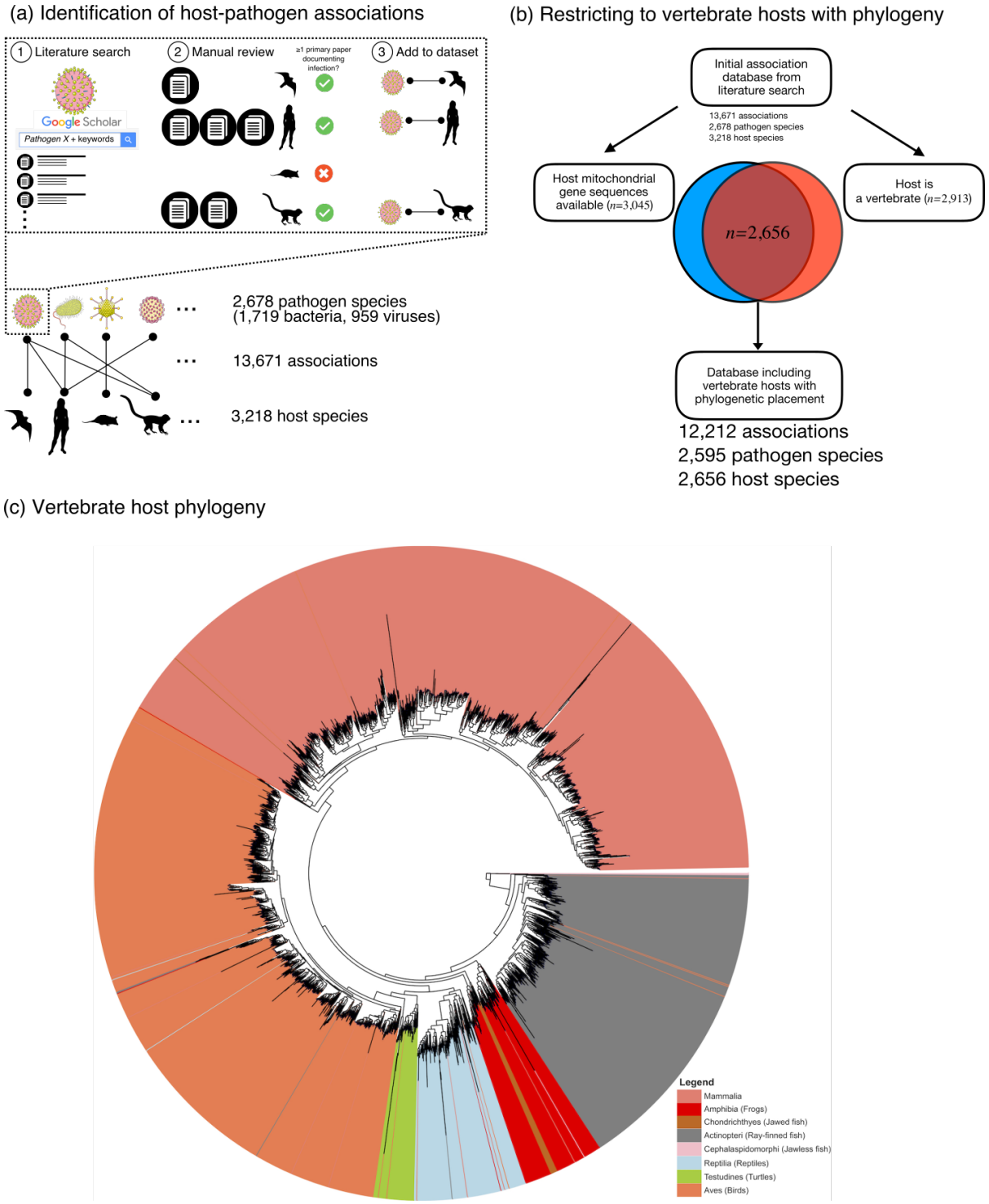


Figure 1. Overview of methodology for compiling the dataset. (a) Methodology of literature review (see Methods). (b) Subsetting the database to only associations involving vertebrate host species for which mitochondrial gene sequences could be identified. (c) Vertebrate host phylogeny. *Image credits: Icons made by Maxim Basinski (tick/cross symbols) and Chanut (document icon), from www.flaticon.com. Pathogen images (influenza, bacterium, adenovirus, HIV) from the Bacteriology Virology image set from Servier Medical Art https://smart.servier.com. Host images (falcon, human, possum, monkey) from PhyloPic.*

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

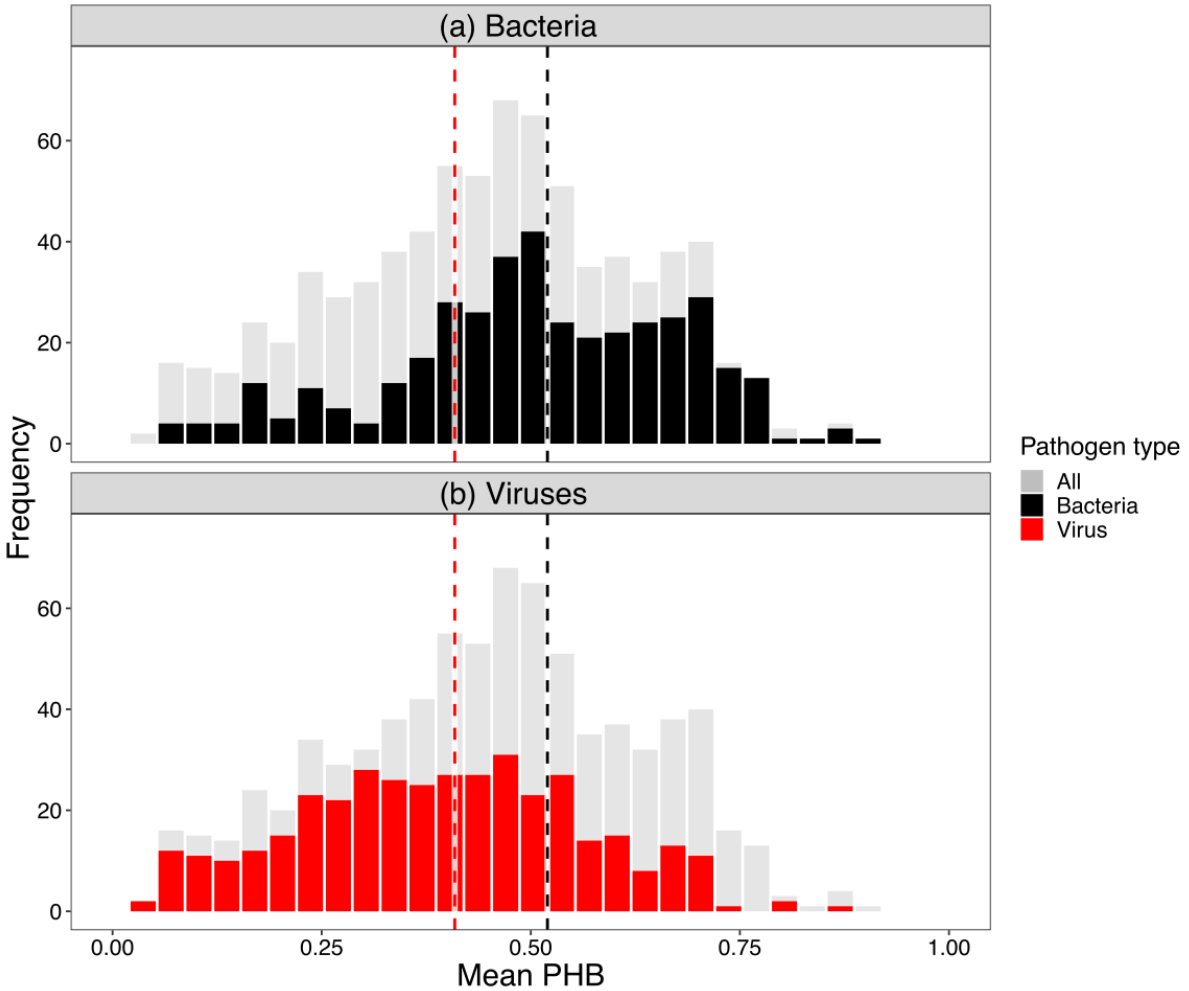


Figure 2. Mean phylogenetic host breadth for multi-host (top panel) bacteria and (bottom panel) viruses. Bacteria and viruses are shown in black and red respectively, with the overall pathogen histogram (both types) shown in grey on both panels to help comparison. On average, multi-host bacteria have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

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

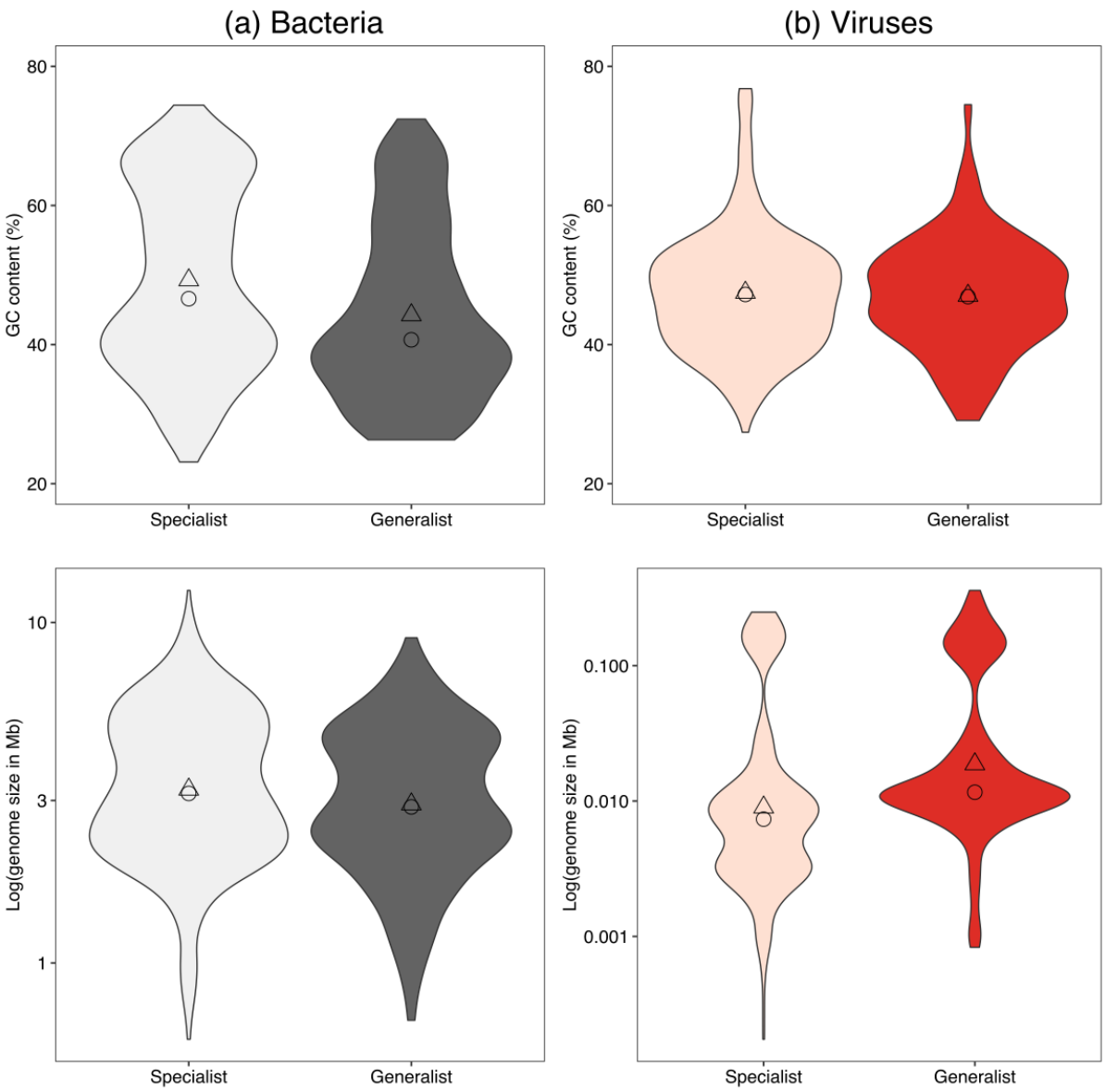


Figure 3. Pathogen genome GC content and size for specialist and generalist pathogens. Note the log-scale for the y-axis in the lower half of the figure. Distributions are shown for specialists (light colours) and generalists (dark colours) with their median (circle) and mean (triangle).

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

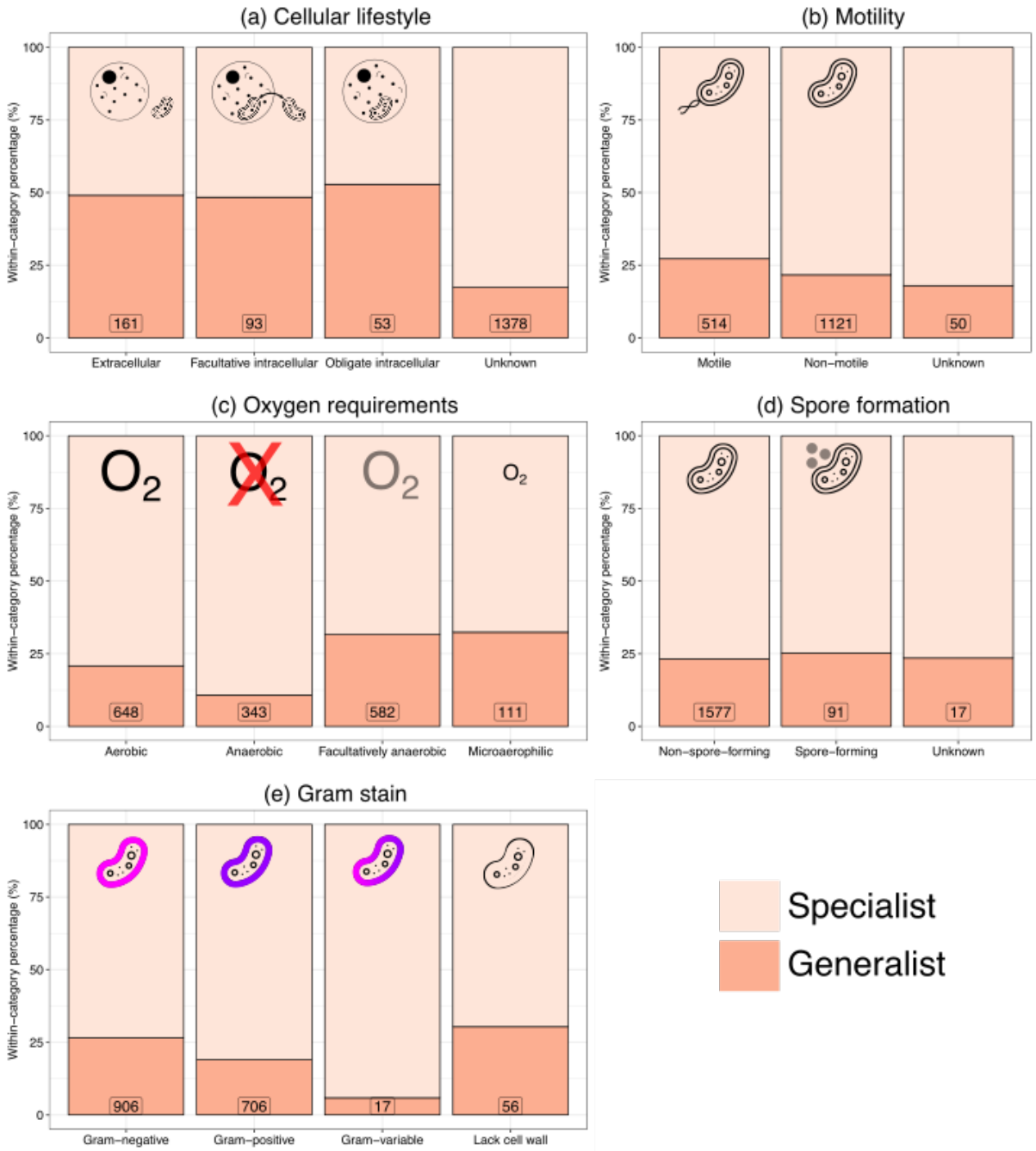


Figure 4. Bacterial lifestyle factors and pathogen range. Proportion of specialists (light pink) and generalists (dark pink, PHB>0) for different categories of bacterial lifestyle: (a) cellular lifestyle, (b) motility, (c) oxygen requirements, (d) spore formation, and (e) Gram stain. 'Unknown' can also mean 'not applicable'.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

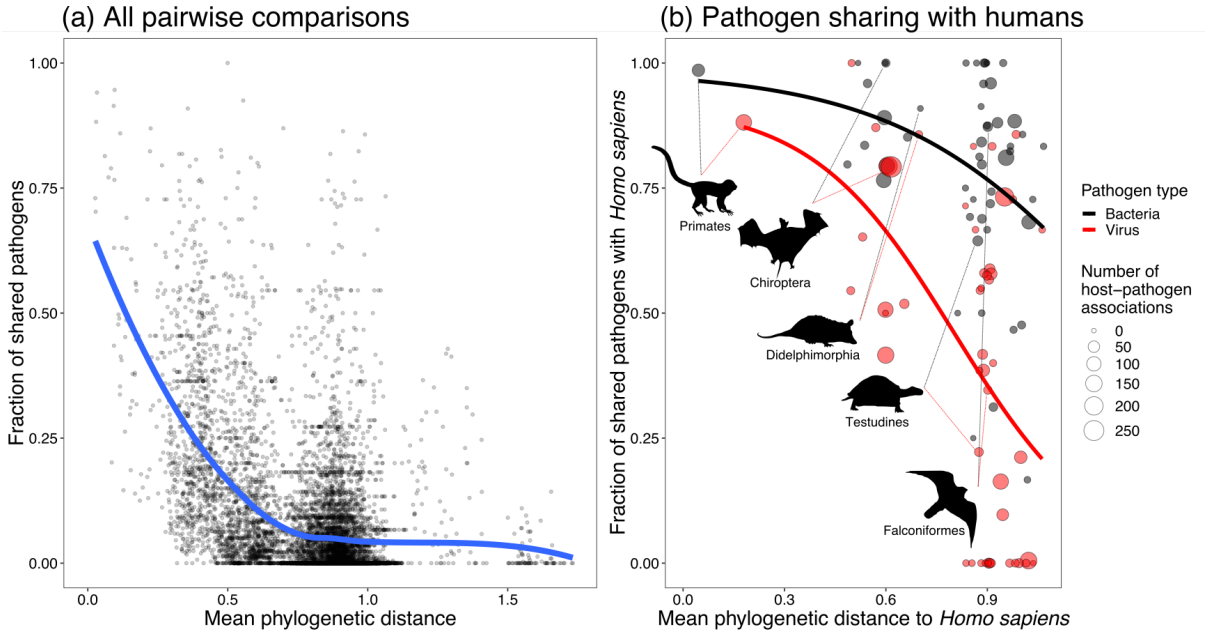


Figure 5. The fraction of shared pathogens between hosts decreases with inter-host phylogenetic distance. (a) All pairwise comparisons between host orders. The blue line shows a smoothed average fit, produced with 'loess'. Only comparisons between host orders with at least 10 host-pathogen associations in the database are shown. (b) Shared pathogens between different host orders and *Homo sapiens* (as a fraction of total pathogens infecting a given 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. Only host orders with at least ten total pathogen-host associations are included. Four illustrative orders are indicated with images.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

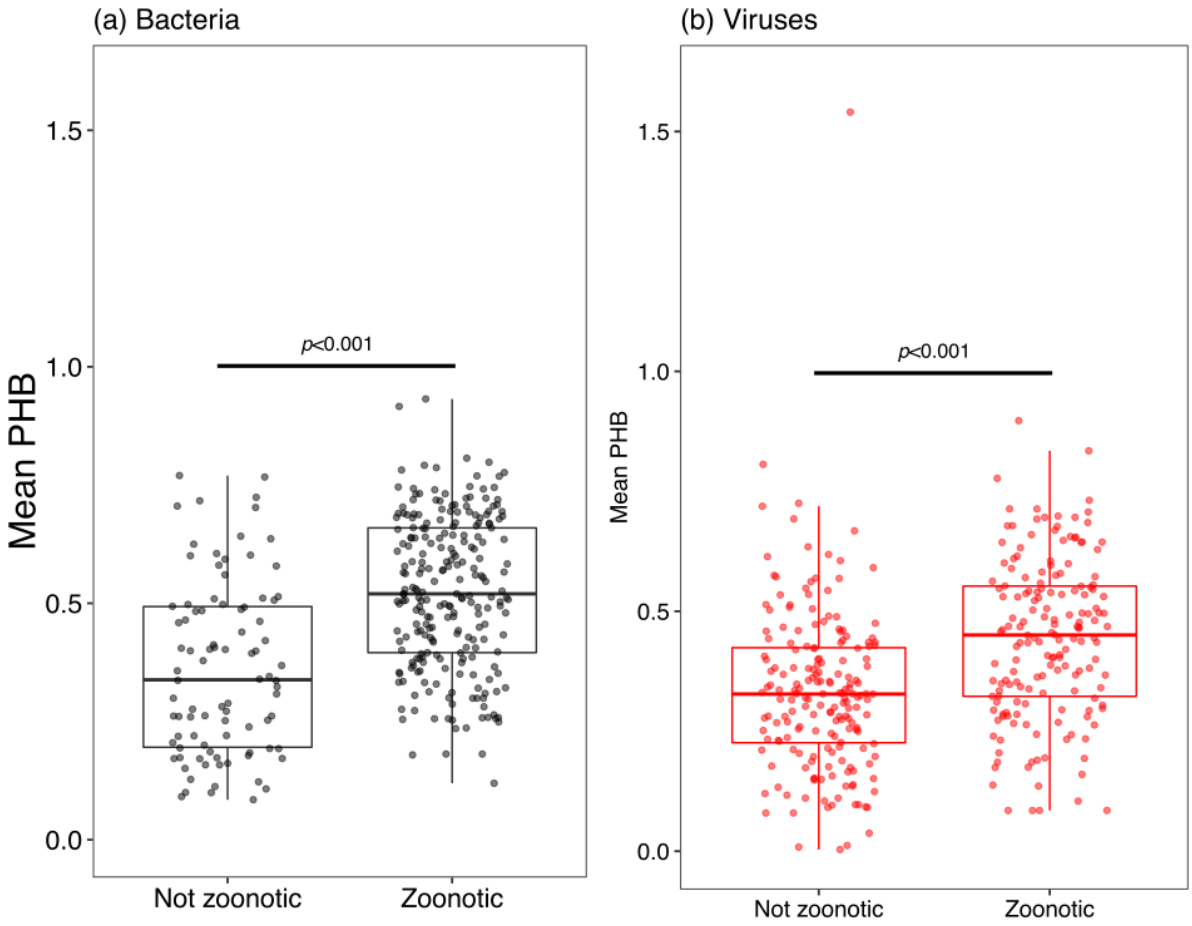


Figure 6. Mean PHB is greater for zoonotic pathogens after excluding human hosts. (a) Bacteria. (b) Viruses. The comparison of zoonotic vs. non-zoonotic pathogens (mean values: bacteria, 0.52 vs. 0.36; viruses, 0.44 vs. 0.33) is a Wilcoxon rank sum test. This plot excludes pathogens that are specialists after excluding human hosts i.e. have only one other vertebrate host. Including these pathogens with a mean PHB of zero did not change the conclusion that zoonotic pathogens had a greater mean PHB (Bacteria: 0.26 vs. 0.03, $p<0.001$; Viruses: 0.31 vs. 0.09, $p<0.001$).

Table 1. Previous studies of the host range of pathogens.

Study	Database details	Methods	Pathogen types	Host range classification	Main conclusions
Taylor et al., 2001 (1)	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 (6)	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 (2)	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 (9)	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 (10)	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.
Han et al., 2016 (51)	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 (11)	586 pathogens across 754 mammal species (2,805 host-pathogen associations)	Literaure 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.

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. - Proportion of shared pathogens between hosts decreases with phylogenetic distance. - Zoonotic pathogens infect more non-human hosts than non-zoonotic.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

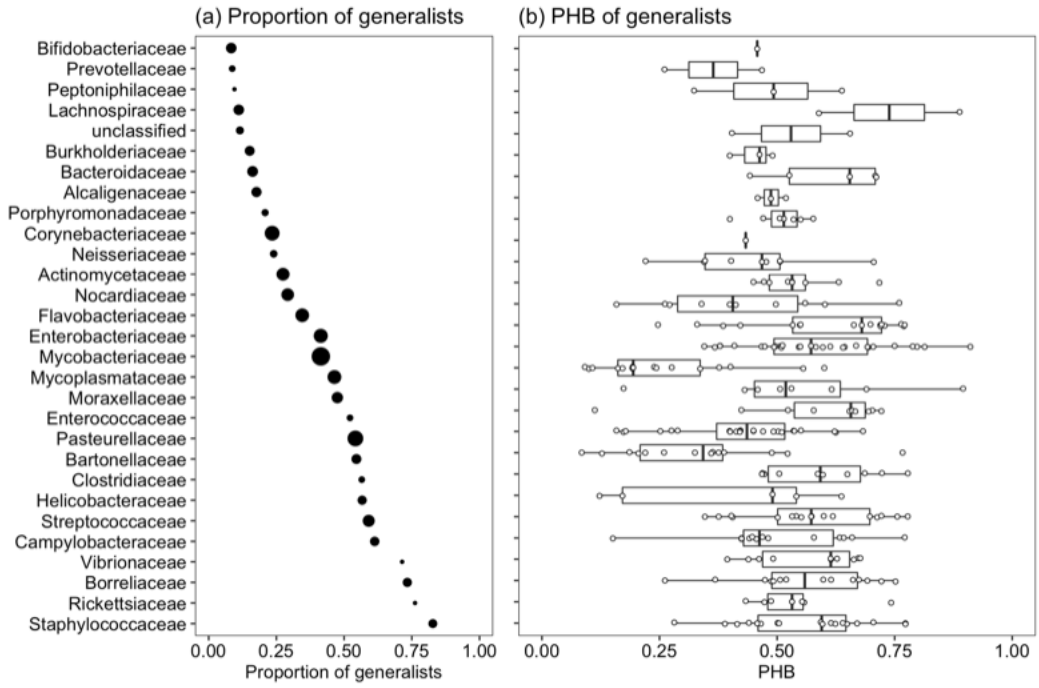
Table 2. Summary of pathogen niche and specificity types. A 'specialist' pathogen infects only a single host, a 'generalist' more than one. Generalists are categorised according to whether their hosts are within the same family (e.g. Bovidae), order (e.g. Artiodactyla), or across orders. Percentages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

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

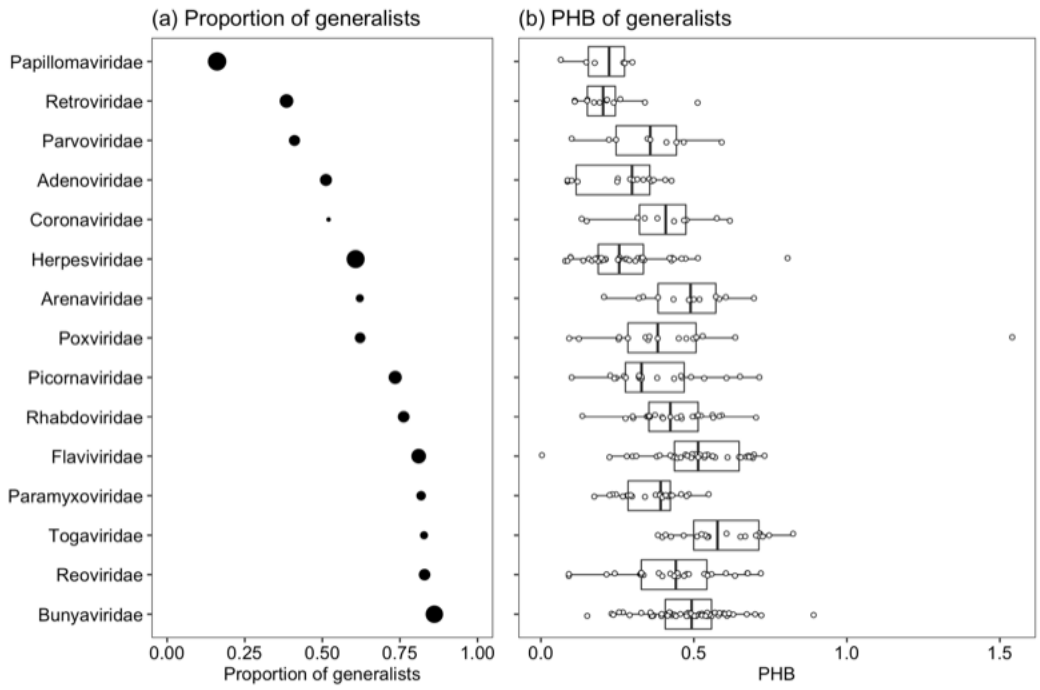
Supplementary Material

Supplementary Text 1. Supplementary analysis notebook. A notebook written in Rmarkdown which reproduces all analysis and figures, as well as additional analysis. Available at figshare (doi: 10.6084/m9.figshare.8262779).

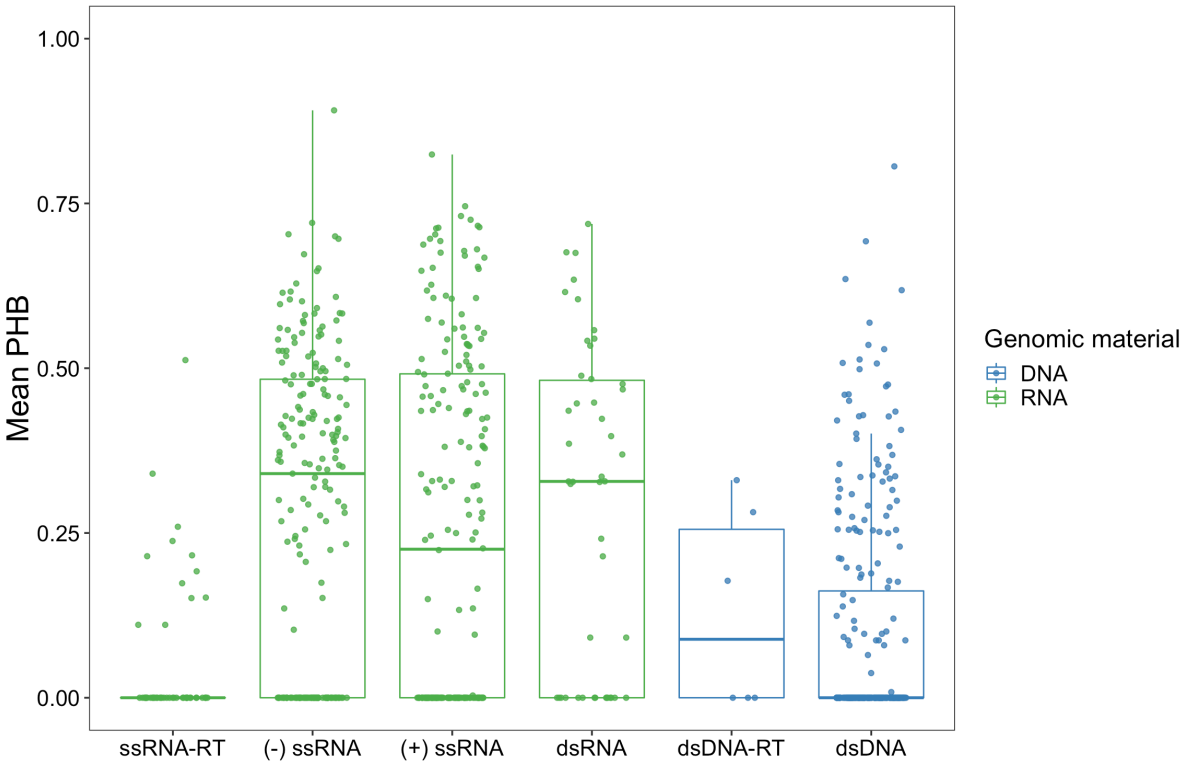


Supplementary Figure 1. (a) Proportion of generalists and (b) PHB of generalist pathogens by bacterial family. Only bacterial families with >20 pathogen species in the association database are shown. Families are ordered by the proportion of generalists. There is no clear association between the proportion of generalists within a family and how wide-ranging those generalists are.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

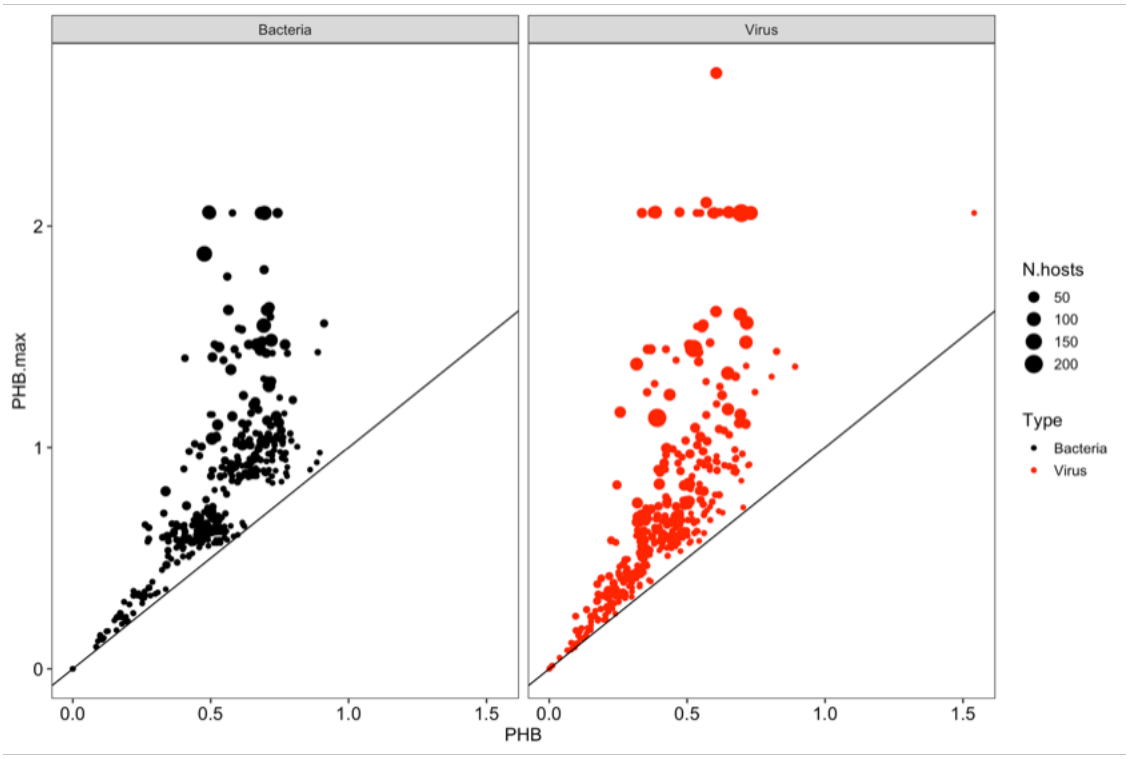


Supplementary Figure 2. (a) Proportion of generalists and (b) PHB of generalist pathogens by viral family. Only viral families with >20 pathogen species in the association database are shown. Families are ordered by the proportion of generalists.



Supplementary Figure 3. Viral genome type is associated with host range. RNA viruses have a greater median mean PHB than DNA viruses. Subgroups shown are the Baltimore classification.

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Supplementary Figure 4. Correlation of mean and maximum PHB for pathogens. Correlation is shown for bacteria (left) and viruses (right).

bioRxiv preprint doi: https://doi.org/10.1101/670315; this version posted June 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Table 1. Vector-borne pathogens are more likely to be generalists. Number of bacteria and viruses with known invertebrate vectors.

Viruses	Specialist	Generalist
Not vector-borne	469	271
Vector-borne	50	117
Bacteria		
Not vector-borne	1240	337
Vector-borne	53	52

Supplementary Table 2. Viruses with an RNA genome and larger genome size have a greater host range. Having an RNA genome and a larger genome were both significantly associated with greater mean PHB ($p<0.001$ for both variables) with a non-significant interaction between them ($p=0.36$). Interestingly, genome size was not significantly associated with greater PHB in a univariate model (Supplementary Text 1).

	Coefficient (s.e.)	*p*
Intercept	0.048 (0.015)	0.002
RNA genome	0.188 (0.031)	<0.001
Genome size	0.661 (0.148)	<0.001
Interaction	1.742 (1.919)	0.365

Supplementary Table 3. Bacterial motility and cellular lifestyle are not associated with greater host range. Combining motility and cellular proliferation in a linear model suggests that neither variable is associated with greater mean PHB. Univariate linear models and a linear model with an interaction term (Supplementary Text 1) give the same conclusion.

		Coefficient (s.e.)	*p*
Intercept		0.264 (0.028)	<0.001
Cellular lifestyle	Facultative intracellular	-0.042 (0.038)	0.278
	Obligate intracellular	-0.004 (0.051)	0.933
Motility	Not-motile	0.025 (0.037)	0.502