PARASITE, HOST, AND ENVIRONMENTAL TRAITS PREDICT THE ZOONOTIC RISK OF PROTOZOAN PARASITES

by

JOY PURNI CHRISTOPHER

(Under Direction of John M. Drake)

ABSTRACT

Zoonotic diseases caused by protozoan pathogens contribute heavily to the global burden of disease. Novel protozoa species which have emerged from wildlife to humans in the recent decades have proven difficult to control. Our ability to anticipate and prevent future emerging disease threats relies on identifying characteristics of zoonotic pathogens and targeting surveillance efforts accordingly. While the traits of zoonotic viruses are well-studies, protozoa have received limited attention. We compiled a dataset of parasites from wild mammal hosts which incorporates both parasite and host traits. Our machine learning model distinguished zoonotic from non-zoonotic protozoa with 85% accuracy. We found that generalist protozoa were most likely to be zoonotic. We ranked the zoonotic potential of protozoa currently not known to be zoonotic to identify parasitic protozoa species of wild mammals which are most likely to be undiscovered sources of current or future zoonoses, identifying them as priority targets for surveillance.

INDEX WORDS: Emerging infectious diseases, protozoa, zoonotic disease, disease ecology, machine learning, macroecology, parasites

PARASITE, HOST, AND ENVIRONMENTAL TRAITS PREDICT THE ZOONOTIC RISK OF PROTOZOAN PARASITES

by

JOY PURNI CHRISTOPHER

B.S., Calvin University, 2016

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2021

PARASITE, HOST, AND ENVIRONMENTAL TRAITS PREDICT THE ZOONOTIC RISK

OF PROTOZOAN PARASITES

by

JOY PURNI CHRISTOPHER

Major Professor: John M. Drake
Committee: Sonia Altizer
Elizabeth G. King

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

For Sarita.

ACKNOWLEDGEMENTS

I would like to express my gratitude to mentors, collaborators, and peers, and countless others who have made this work possible.

I am fortunate to have benefitted from the mentorship of several talented scientists. First, I would like to thank my advisor, John Drake, for introducing me to the world of disease ecology and encouraging me to pursue new ideas and exciting questions. Your passion for science and dedication to conducting high-quality research is something I always admire. To my other committee members, Sonia Altizer and Lizzie King, I thank you for guiding me with enthusiasm at every step of the way. Outside of my committee, I want to especially thank Barbara Han, Michelle Evans, and Claire Teitelbaum for their exceptional mentorship and assistance with this project. Thank you for thinking through things with me, checking my code, reading drafts over and over, and looking out for my wellbeing. I am infinitely grateful to Barbara, who is kind of collaborator and mentor I aspire to be. You are one in a million, and I feel lucky to have met you and worked with you. Michelle is my peer mentor, collaborator, and dear friend, who has constantly gone above and beyond to support, motivate, and advocate for me even though it's not her job. She is one of the kindest and smartest people I know. Thank you for keeping me accountable. You have challenged me to learn about and address injustices in academia and broader society. I hope you know that you have been instrumental in helping me grow into a better scientist and person. I would also like to acknowledge the guidance I received from Vanessa Ezenwa, Nicole Gottdenker, Craig Osenberg, and Amira Roess. I am grateful to this team of mentors for

being constant sources of knowledge, wisdom, and encouragement. Each of you has readily made time to lend me a listening year as I navigated difficult situations, both personal and professional.

I have had the privilege of being a member of the Drake lab over the last few years, and enjoyed interacting and developing meaning relationships with many of you. To Andrea, Drew, Reni, Tad, Michelle, Paige, Robbie, Nikki, Dom, Anna, Eamon, Eric, Spencer, Tierney, Trippe, Cecilia, Kyle, Jo, Emmanuel, Mallory, Kailah – thank you for shaping a lab culture that felt like home me, a place where I could belong. In my biased opinion, I think we're one of the coolest and most fun labs in the Odum School. I have gleaned so much from you and will carry this with me going forward.

Graduate school can be grueling, and I could never have made it through without the support of the Odum School graduate student community. To Carolyn, Denzell, and Carol – you are my ride-or-die, and I don't know what I would do without you. I want to thank other members of my cohort and fellow IDEAS students – especially Reed, Zach, Nikki, Deven, Mike, Kaylee, and Kate Sabey – for being comrades in arms. We got through a ridiculous amount of intense coursework together. I will fondly remember our late-night study sessions which often turned into all-nighters spent on campus, and out spontaneous social outings, which were always a blast. Thank you for carving our space for venting sessions where we could freely air our grievances, as well as encourage and motivate each other, and celebrate each other's successes. There are so many other graduate students – Claire, Robbie, Supraja, Laura, Talia, Cecilia, Akanksha, Annakate, and Caitlin, to name a few – who have supported me academically and socially. You will always have a special place in my heart.

I am also grateful to the broader Odum School community. Many staff members have supported me throughout the years. I thank Ford Ballantyne, Katherine Adams, Craig Osenberg, and Julie Escobedo-Gunby for serving as graduate program coordinators and mentors. I would be lost without IT support from Brian Perkins and Tyler Ingram. It was always a joy to work with Beth Gavrilles. I appreciated the opportunity to serve on several committees, especially the DEI committee.

I want to thank my therapist for their support over the last four years, which has helped me persevere through times of difficulty. I would also like to acknowledge the Trevor Project and National Suicide Prevention Lifeline for their important work.

I'm forever grateful to Laura and Robbie for being an amazing friends and roommates. You have been a pillar of support, and I couldn't ask for better people to be cooped up with during a pandemic. Laura, you have been a wonderful companion throughout the last three years. Thank you for your patience and kindness. Shout out to the canine companions I've had the privilege of sharing a home with, and all those that have provided me with many cuddles and cuteness when I needed them. Izzy, CJ, Coco, Catalina, Lula, Skip, Wahala, Cray, and Xena – you are the goodest.

Finally, I want to thank my parents for always believing in me, even from the other side of the world. There are no words that can convey how much your care and love has meant to me. Thank you for making me who I am. I love you always.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Emerging zoonoses

Emerging infectious diseases (EIDs), which are “infections that have newly appeared in a population or have existed previously but are rapidly increasing in incidence or geographic range” (Morse 1995), pose a serious threat to global public health. The number of EIDs has increased over time, and over 60% of them are zoonotic (Jones *et al.* 2008; Karesh *et al.* 2012), meaning that they are transmitted from non-human animals to humans. Zoonotic diseases are a growing concern because they can emerge quickly unpredictably anywhere, and can spread rapidly around the globe. EIDs such as HIV and COVID-19 continue to have enormous social, economic, health costs (Bender *et al.* 2006; Martins *et al.* 2015; Huber *et al.* 2018). Our ability to anticipate future emerging disease threats relies on identifying risk factors that lead to the emergence of zoonotic infectious diseases.

In the past decade, substantial research effort has been put towards understanding how and why zoonotic diseases emerge in human populations (Gortazar *et al.* 2014; Salkeld *et al.* 2016; Allen *et al.* 2017), with several key findings that can be used to guide surveillance efforts to prevent and control zoonotic emergence. First, zoonotic diseases are caused by pathogens from all major parasite taxa; an estimated 80% of viruses, 50% of bacteria, 40% of fungi, 70% of protozoa, and 95% of helminths that infect humans are zoonotic (Morse *et al.* 2012). This finding suggests that surveillance

efforts cannot target a single parasite group, but must consider diverse parasites. In addition, mammals are the main reservoirs for zoonotic diseases, accounting for roughly 80% infections of that spill over into humans (Taylor *et al.* 2001). Among pathogens shared between humans and non-human mammals, ungulates, rodents, carnivores, and primates are common hosts (Cleaveland *et al.* 2001). However, many factors that could lead to disease emergence remain unexplored, including details of parasite and environmental traits that might predict spillover. Given the increasing rate of disease emergence, it is in our interest to understand the risk factors for emergence, so that we can anticipate emerging zoonotic disease threats, respond proactively, and prevent emergence and spread.

Ecological theory and empirical tools are useful for studying host and parasite species and for understanding the traits of zoonotic parasites. We have compiled databases of host-parasite interactions across the globe, which can be leveraged to quantify large-scale patterns of host-parasite ecology. These databases can be analyzed using recent advances in machine learning methods, which address problems posed by more traditional methods, such as correlations between variables, non-linear relationships, phylogenetic non-independence of trait data, variation in sampling effort, and incomplete or missing data. Machine learning algorithms such as boosted regression trees are robust to these issues, and are gaining popularity in disease ecology research (Han *et al.* 2020; Pandit & Han 2020). These tools and datasets provide an opportunity to further investigate the ecology of EIDs.

Thesis Objectives

Prior research has investigated a variety of extrinsic factors contributing to zoonotic disease emergence – such as human demographics, the industrialization of food production, globalization, international travel and commerce, land use, microbial adaptation, and changes and breakdown in public health systems (Altizer et al. 2013; Schmeller et al. 2020; Plowright et al. 2021). Understanding these relationships has improved our ability to anticipate emerging disease threats. In addition to extrinsic factors, the species traits can influence the potential for zoonotic emergence. For example, host species traits can determine reservoir potential (Han et al. 2015, 2019; Plourde et al. 2017 Luis et al. 2015; Olival et al. 2017). Likewise, parasite species traits can determine their zoonotic potential (Flanagan et al. 2012; Johnson et al. 2015a; Evans et al. 2018; Walker et al. 2018; Park 2019). Therefore, compiling data on the traits of known zoonotic pathogens can help identify pathogens that are likely to be zoonotic, based on their trait similarity to known zoonotic pathogens.

Significant research efforts have been put towards understanding the traits of zoonotic viruses and their hosts, which are the main source of emerging zoonoses. However, other major parasite taxa such bacteria, helminths, protozoa, fungi, and prions are relatively understudied. The traits of zoonotic protozoa in particular have received limited attention, despite contributing heavily to the global burden of disease. Predicting which protozoan parasites of wild mammals are likely to be undiscovered sources of future zoonoses is an important strategy for countering emerging disease threats, because this knowledge can guide policies and inform priorities for wildlife disease surveillance.

Aims

I. Identify traits of parasite species that distinguish zoonotic protozoa from non-zoonotic protozoa.

II. Predict which protozoa species are most likely to be undiscovered sources of future zoonoses.

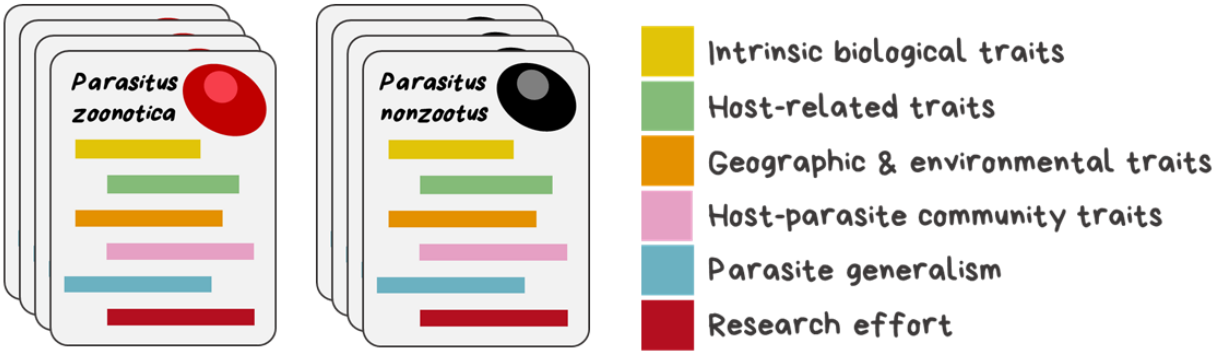
Questions

1) Which traits of protozoa of wild mammals (ungulates, primates, and carnivores) are significant predictors of zoonotic status?
2) Are intrinsic parasite characteristics, host-related characteristics, host-parasite network properties, environmental traits, community traits, parasite generalism, or research efforts more important for predicting zoonotic potential?
3) Which protozoa species that are not currently known to be zoonotic are predicted to have high zoonotic potential?

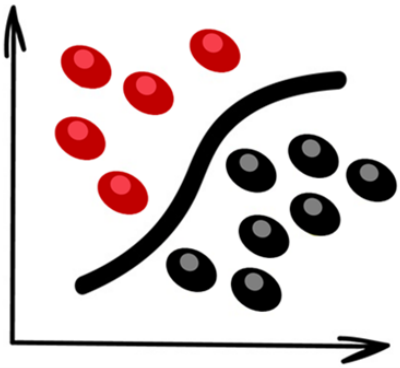
Strategy

We used records of host-parasite associations from the Global Mammal Parasite Database (GMPD) version 2.0 (Nunn & Altizer 2005; Stephens *et al.* 2017), and assigned zoonotic status codes to each parasite species. We then collected traits of protozoa species in our dataset and trained a statistical model to identify traits that distinguish between zoonotic and non-zoonotic protozoa. We then used the model to calculate zoonotic risk scores for each parasite and identified which non-zoonotic protozoa posed a high risk of becoming zoonotic. An overview of this methodological approach is illustrated in Figure 1.1.

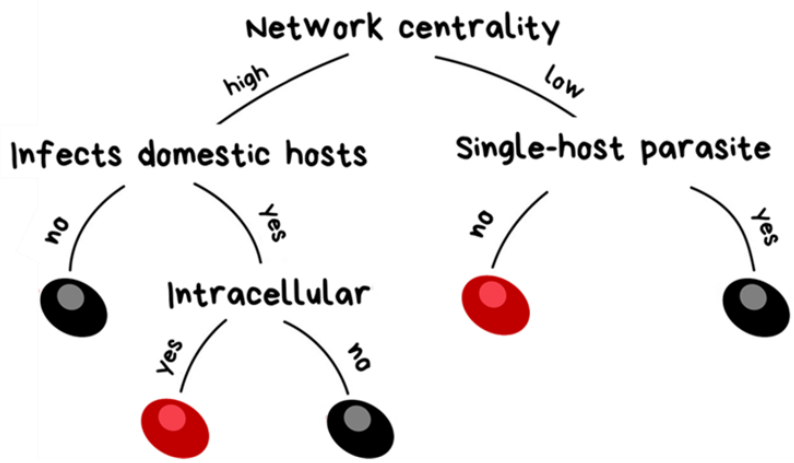
1. Collect six types of species-level traits to create a trait profile for each protozoan parasite in the dataset.



2. Train a machine learning model to distinguish between zoonotic and non-zoonotic protozoa.

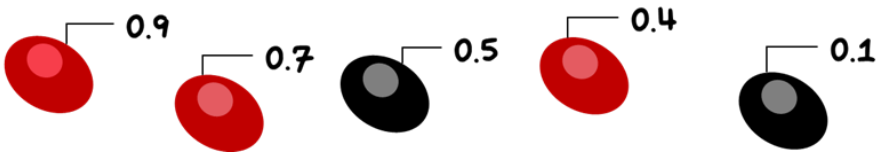


3. Identify traits that were most important for accurately predicting zoonotic status.



Binary classification with Boosted Regression Trees (BRT)

4. Assign risk scores to parasites based on traits the model identified as important predictors of zoonotic status.



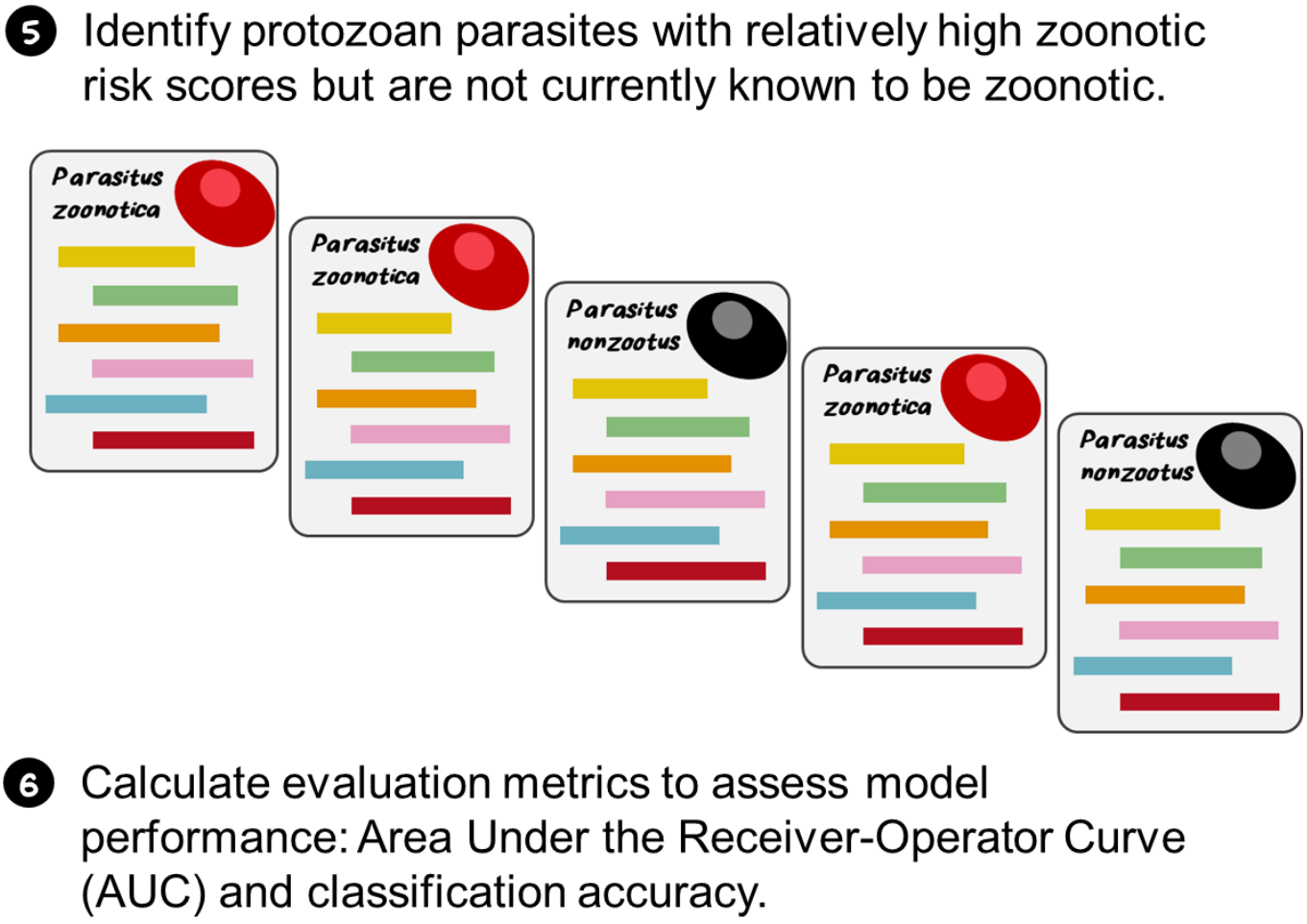


Figure 1.1. Conceptual diagram of the methodological approach used in this study.

Steps 1-3 correspond to Aim I and steps 4-6 correspond to Aim II.

CHAPTER 2

PARASITE, HOST, AND ENVIRONMENTAL TRAITS PREDICT THE ZOONOTIC RISK OF PROTOZOAN PARASITES[1]

[1] Venkatachalam-Vaz, J., Han, B.A. & Drake, J.M. (2021). Parasite, host, and environmental traits predict the zoonotic risk of protozoan parasites.

To be submitted to *International Journal for Parisitology*

Abstract

Protozoan zoonoses, such Chagas disease and leishmaniasis, remain endemic in large parts of the world, exacerbating social inequity and contributing heavily to the global burden of infectious disease. Novel protozoa species which have emerged from wildlife to humans in the recent decades (e.g., *Plasmodium knowlesi*, a causal agent of malaria) have proven difficult to control. Our ability to anticipate and prevent future emerging disease threats relies on identifying the characteristics of zoonotic pathogens and targeting surveillance efforts accordingly. While several studies have profiled the traits of zoonotic viruses, protozoa have received limited attention. We compiled a dataset of protozoa species which incorporates both parasite and host traits, including information on community structure and importance within a host-parasite bipartite network. Using a machine learning algorithm, extreme gradient boosting, we distinguished zoonotic from non-zoonotic protozoa with 85% accuracy. Our model found that traits of generalist protozoa (e.g., broad tissue tropism, high network centrality, multiple transmission modes) were most useful for predicting zoonotic status, compared to intrinsic biological traits (e.g., morphology), environmental traits (e.g., temperature), or host-related traits (e.g., life history). Here we report parasitic protozoa species of wild mammals which are most likely to be undiscovered sources of current or future zoonoses, identifying them as priority targets for surveillance.

Introduction

As novel pathogens emerge into human populations, they pose a serious threat to global health. The majority of emerging pathogens are zoonotic, i.e., are transmitted from non-human animals to humans (Woolhouse *et al.* 2001; Jones *et al.* 2008). Research on a variety of extrinsic factors that contribute to zoonotic disease emergence – such as climate change, land use, and biodiversity – has improved our ability to anticipate emerging disease threats (Altizer *et al.* 2013; Schmeller *et al.* 2020; Plowright *et al.* 2021). Recent studies have also found that the traits of hosts, pathogens, and host-parasite networks can be important factors for predicting zoonotic risk (Luis *et al.* 2015; Olival *et al.* 2017). Literature in this area focuses on identifying host traits, which determine potential reservoir species (Han *et al.* 2015, 2019; Plourde *et al.* 2017), and the characteristics of parasite species – especially viruses – which determine their zoonotic potential (Flanagan *et al.* 2012; Johnson *et al.* 2015; Evans *et al.* 2018; Walker *et al.* 2018; Park 2019).

While significant research efforts have been put towards understanding the traits of viral zoonotic pathogens, which have been responsible for a large proportion of emerging infections in recent years (Nii-Trebi 2017), other major parasite groups such bacteria, helminths, protozoa, fungi, and prions have been relatively understudied. Because the ecology and epidemiology of viruses differs from that of other major parasite taxa, traits that are important predictors of the zoonotic potential of viruses cannot be expected to apply to other groups.

Zoonotic protozoa in particular have received limited attention, despite contributing heavily to the global burden of disease. Malaria, caused by protozoan

parasites in the genus *Plasmodium*, was responsible for over 230 million cases, 600,000 deaths, and 46,000 disability-adjusted life years (DALYs) in 2019 alone (Institute for Health Metrics and Evaluation (IHME) 2019). Though malaria is an ancient disease that has infected humans for tens of thousands of years, novel malaria species have emerged from wildlife to humans in the last few decades, such as *Plasmodium knowlesi* (Singh *et al.* 2004; Cox-Singh *et al.* 2008; Ahmed & Cox-Singh 2015). The emergence of *P. knowlesi* poses new challenges for malaria-control efforts because the existence of a wildlife reservoir makes eradication more difficult (Brock *et al.* 2016). Other diseases caused by zoonotic protozoa that spill over from wild mammals, such as toxoplasmosis (*Toxoplasma gondii*), visceral leishmaniasis (*Leishmania infantum*), and Chagas disease (*Trypanosoma cruzi*), have proven difficult to control and remain endemic in large parts of the world (Torgerson & Macpherson 2011; Pisarski 2019). In addition to contributing significantly to global infectious disease burden in terms of cases, deaths, and DALYs, these diseases cause significant economic losses and exacerbate social inequalities (Herricks *et al.* 2017). Identifying traits that are associated with zoonotic protozoa can help us anticipate and prevent future disease emergence.

Zoonotic diseases are typically considered to be those caused by the directional transmission of a zoonotic parasite from an animal reservoir to a human, either through direct interspecies contact, or indirectly through biting arthropod vectors or through contaminated food or water. However, there are numerous human infectious diseases caused by parasites for which this directional transmission (arising directly from an animal reservoir) has not yet been firmly established. These parasites may be found in animals as well as humans, both hosts having acquired the pathogen from a common

source (e.g., environmental). Traits associated with known zoonotic parasites (e.g., life history strategy, morphological traits, or infection characteristics) can be used to predict which currently non-zoonotic parasites pose a high risk of spilling over into humans from animal reservoirs.

Though other studies that examined host traits or parasite traits separately have yielded useful predictions and identified targets for surveillance, combining both categories of traits could result in more accurate predictions by incorporating the full ecology of disease transmission. Particular combinations of pathogen and host characteristics determine host range, thus considering them together rather than separately may offer clues to which parasites exhibit true zoonotic transmission from animals and which non-zoonotic parasites merely happen to infect animals and humans. Here, we incorporate both sets of traits, as well as information on the network and community structure of host-parasite interactions, to discriminate zoonotic from non-zoonotic protozoa.

We compiled a species-level dataset of protozoan parasites of wild mammals (primates, carnivores, and ungulates), classified each parasite as “zoonotic” or “non-zoonotic”, and collected traits to be used as variables which we hypothesized could be useful for predicting class membership (i.e., zoonotic status). We defined as “zoonotic” only those parasites with evidence of directional transmission (animal sources causing infection and disease in humans) and considered these distinct from parasites that are merely shared with humans for which the evidence for direction transmission is not yet firmly established. We then trained a machine learning model to distinguish zoonotic from non-zoonotic protozoa based on the trait variables in our dataset. This model

identified trait variables that were most useful for predicting zoonotic status and provided risk scores for each protozoa species signifying that parasite's probability of being zoonotic. We then used those risk scores to identify particular non-zoonotic protozoan parasites with the highest predicted risk of zoonotic transmission (directional spillover from animals into human populations) and suggest priorities for surveillance and control in wild animal populations.

Theoretical approach

To better understand which traits are most important for zoonotic emergence of protozoan parasites, we compared the relative importance of five trait categories: (i) the degree of parasite specialism or generalism, (ii) intrinsic biological traits, (iii) host-parasite network and community characteristics, (iv) geographic and environmental traits, and (v) host-related traits. Below, we describe how and why these traits might be important for predicting zoonotic potential.

Parasite traits, including generalism and intrinsic biological traits, are important to a parasite's ability to infect host species, and thus might be a crucial contributor to zoonotic potential (Leggett *et al.* 2013; Park *et al.* 2018). We hypothesized that parasites with a high degree of generalism (measured by the number and diversity of host species, tissue types, transmission modes, and environments they are associated with) are more likely to be zoonotic than specialist parasites because they have a broader repertoire for overcoming species barriers. We also hypothesized that zoonotic protozoa would be likely to have non-close transmission modes (foodborne, waterborne, or vector-borne) because close contact between humans and wildlife is uncommon relative to livestock or companion animals (Craft 2015; Fong 2017). Finally, we

predicted that taxonomic group would be predictive of zoonotic potential because closely related species are more likely to share traits that enhance cross-species transmission potential than unrelated species.

In addition to these parasite-specific traits, characteristics of the host-parasite community could be important for parasite zoonotic potential. In a bipartite network connecting parasites and their associated hosts (Fig. 2.1), we hypothesize that parasites with higher node importance are more likely to be zoonotic than those on the network periphery because well-connected parasites in the host-parasite community might be able to infect a greater variety of hosts, including humans. Similarly, non-zoonotic protozoa that share hosts with many zoonotic protozoa might have high zoonotic potential, because their hosts are successful transmitters of zoonoses. Lastly, the “bridge host hypothesis” postulates that domestic animals can facilitate transmission from wildlife to humans (Wolfe *et al.* 2007; Caron *et al.* 2015), so we expected that protozoa that infect domestic animals would have higher zoonotic potential via more frequent, direct exposure to humans via livestock.

Additionally, we considered the environmental traits or climatic conditions associated with the geographic ranges of zoonotic and non-zoonotic protozoa, to identify variables which contribute to spillover. This included traits such as mean human population density, temperature, and precipitation, and biogeographic region at parasite locations. Previous research has shown that climate and urbanization are important drivers for emerging zoonotic protozoa such as *Plasmodium knowlesi* (Fornace *et al.* 2019).

Finally, we included traits representing host ecology such as mean social group size, trophic level, and pace of life, to investigate differences in hosts-related traits of zoonotic and non-zoonotic protozoa.

Our model was able to distinguish zoonotic status among protozoa 82% of the time and identified parasite generalism as the most important trait category for predicting zoonotic status. We also identify five protozoa not confirmed to be zoonotic that had a high estimated likelihood of zoonotic transmission from wild mammals to humans: *Neospora caninum*, *Entamoeba histolytica, Entamoeba dispar, Leishmania donovani,* and *Leishmania braziliensis*.

Methods

Dataset Assembly

Global Mammal Parasite Database

We obtained records of host-parasite associations from the Global Mammal Parasite Database (GMPD) version 2.0 (Nunn & Altizer 2005; Stephens *et al.* 2017), which contains records of over 24,000 observations of interactions between wild mammalian hosts (ungulates, carnivores, and primates) and parasites (viruses, bacteria, protozoa, helminths, fungi, prions, and arthropods). We scored the zoonotic potential of 1,453 microparasites from the GMPD, using a standardized protocol detailed in Appendix A. Each parasite was assigned one of five *zooscores* – integers ranging from -1 to 3. These are defined in Figure 2.1. For the purposes of our study, the five *zooscores* were converted into binary scores dividing parasites into two classes: known to be zoonotic (1) or not know to be zoonotic (0). The binary zoonotic status codes were used as the response variable in our machine learning model, allowing the algorithm to distinguish between the two classes of parasite. Parasites with a zooscore of -1 or 0 were considered non-zoonotic, while parasites with a zooscore of 1, 2, or 3 were considered zoonotic.

Parasites assigned a *zooscore* of 0 (i.e., those found in both human and non-human vertebrates) were classified as non-zoonotic because there was no established evidence documenting animal-to-human transmission. For example, *Giardia intestinalis* (also known as *Giardia duodenalis* or *Giardia lamblia*) is a common intestinal parasite infecting a variety of mammalian host species. *Giardia* is spread through the fecal-oral route, and infection can be acquired by consumption of infective cysts in contaminated

food or water (Feng & Xiao 2011). Currently, empirical tools available to study the molecular epidemiology of parasites are unable to identify the sources of *G. lamblia* infection in humans and animals, or track the direction of transmission (Ryan & Cacciò 2013). In such cases where zoonotic transmission of shared protozoa has not been confirmed but cannot be ruled out, our model was used to estimate the probability of undiscovered zoonotic transmission.

We assigned zooscores to a total of 1,453 GMPD parasite species. Of these 203 (14%) were in the zoonotic class and 1250 (86%) were in the non-zoonotic class. Of the 1453 zooscored parasite species, 226 were protozoa – 19 (8%) in the zoonotic class and 207 (86%) in the non-zoonotic class. These 226 protozoa species and their 244 associated hosts formed the basis of our dataset. We collected trait data related to these species and their interactions.

We first extracted taxonomic classification and transmission mode variables from the GMPD. Taxonomic traits included the phylum, class, order, and family of each protozoa species. We also extracted transmission mode (TM) data from the GMPD for each protozoa species that had TM information. This was recorded as four binary variables indicating which of the following TMs were used by parasite species in our database: close (transmissible via close non-sexual contact such as grooming, biting, scratching, aerosols), non-close (transmissible via non-close contact such as by fomites or ingestion of food or water contaminated with feces or urine), intermediate (transmitted by intermediate hosts such as snails or crustaceans), and vector (transmissible by biting arthropod vectors). We also calculated the total number of TMs exhibited by each species.

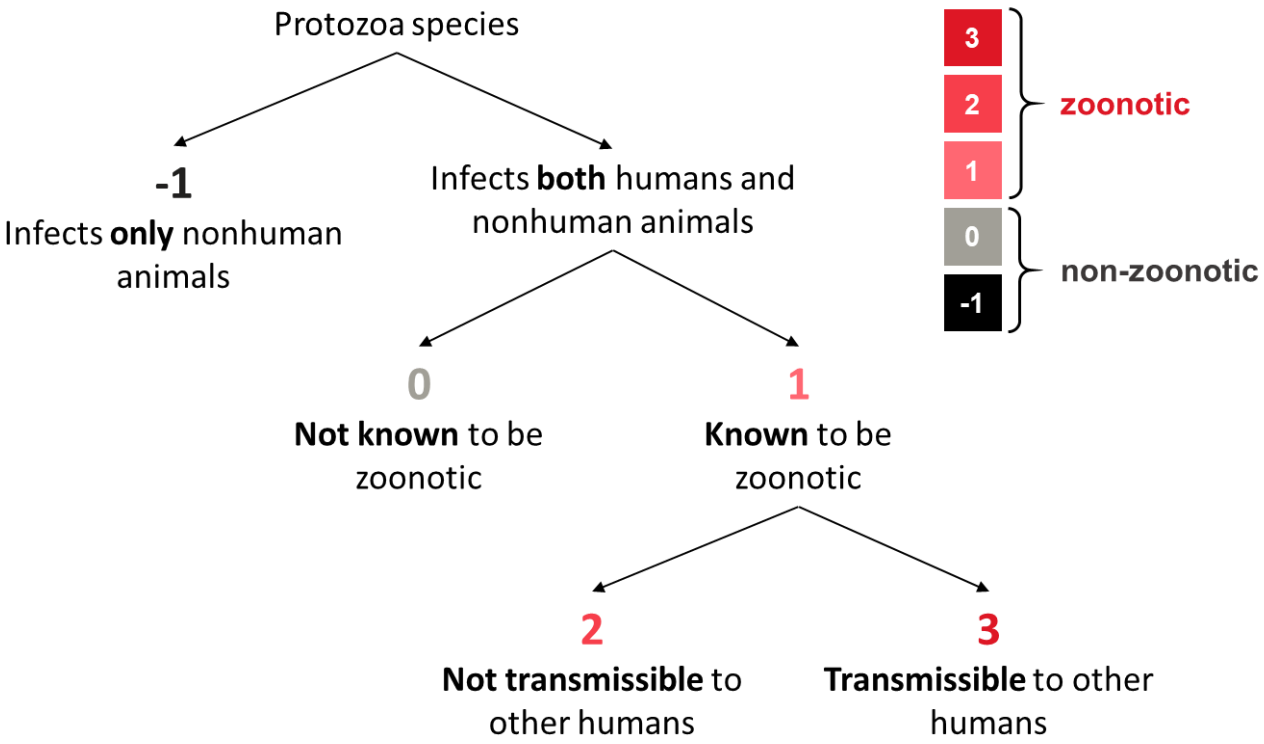


Figure 2.1. Flowchart for assigning zoonotic status codes, or "zooscores" to GMPD parasites. Microparasites with zooscores of 1, 2, or 3 (red) were considered zoonotic, and those with a zooscore of 0 or -1 (grey and black) were considered non-zoonotic according to our criteria.

We added a final variable tallying the number of Web of Science (WOS) citations for each protozoa species in our dataset, as of February 6, 2017. This was used as a proxy for research effort and was included in the model to control for potential sampling biases in data collection.

Primary literature

We searched the primary literature to extract intrinsic traits of each protozoa species in our dataset, which included morphological (flagellate or non-flagellate), life history (cyst stage or no cyst stage), reproductive (sexual or purely asexual), and pathological (intracellular or extracellular parasitism, site of infection) characteristics. We also recorded whether the parasite infected domestic animals (cattle, sheep, goats, horses, dogs, cats, swine, poultry, camels, or other livestock). To calculate the number of anatomical sites that a parasite is capable of infecting, we divided the host body into 13 distinct mammalian organ systems and recorded which organ system(s) each parasite was known to infect. The 13 mammalian organ systems were: muscular, skeletal, circulatory, respiratory, digestive, immune, urinary, nervous, endocrine, reproductive, lymphatic, integumentary, ocular. We included the total number of different organ systems each protozoa species was capable of infecting as another intrinsic variable.

Mammalian host data

Host trait data were extracted primarily from PanTHERIA, (Jones *et al.* 2009), and PHYLACINE (Faurby *et al.* 2018), global species-level databases of all known mammals. From these databases, we extracted traits of a host species which might influence the number and diversity of parasites it is exposed to, such as diet breath,

habitat breadth, geographic range, degree of geographic isolation, and trophic level. We also extracted host population densities, mean human population densities in host habitats, and climatic conditions because we expect these variables to affect transmission dynamics in ways that impact the likelihood of zoonotic spillover. All host-related variables were numerical. To incorporate host species-level traits into our parasite species-level dataset, we calculated the median values across all host species associated with each parasite.

Finally, we obtained values for host geographic ranges and measures of evolutionary distinctiveness (ED), as reported in Dallas *et al.* (2019). ED scores were calculated using the fair proportions method (Isaac *et al.* 2007). We hypothesized that zoonotic parasites would be associated with less evolutionarily distinct hosts.

Ecoregion data

Observations of host-parasite associations in the GMPD are georeferenced, providing coordinates for the majority of records. We used these coordinates to extract ecoregion data from Terrestrial Ecoregions of the World (TEOW) database (Olson *et al.* 2001). TEOW realms were defined as “continental-scale biogeographic regions defined by differences in geologic and climatic history that contain distinct assemblages of plants and animals”. The eight realms were: Australasia, Antarctic, Afrotropic, Indomalaya, Nearctic, Neotropics, and Oceania. We calculated the number of ecoregions for each protozoa species that had georeferenced records (n = 224) as well as its the primary biogeographic region (or realm). The primary biogeographic region assigned to a parasite was the realm it was most frequently recorded in. In the small number of cases where points occurred equally in two biogeographic regions (n=9), we

arbitrarily assigned the species to one of the regions. In most of these cases, the parasite species had only two occurrence records in the GMPD.

Bipartite network analysis

GMPD host-parasite association records were used to construct a bipartite network of all hosts and all parasites in the GMPD (not only protozoa). This was done using the *bipartite* package in R (Dormann *et al.* 2008),which is designed for two-level ecological networks, such as seed-disperser, plant-pollinator, and host-parasite systems. The resulting two-mode interaction matrix consisted of two groups of nodes: a higher-level group (parasite species) and a lower-level group (host species). An edge was drawn between host nodes and parasites nodes for which there was a record of association in the GMPD. Edges were weighted by the number of times each host-parasite association was observed.

We used this bipartite network to calculate several network properties for each node. The *specieslevel* function was used to compute the following three indices for both host and parasite species nodes in the network: (I) betweenness centrality, (II) closeness centrality, and (II) proportional generality. Betweenness centrality measures the number of times a node lies on the shortest path between other nodes. Higher values indicate that a species acts as “bridge” between species in the network. It can be used to identify parasites that share hosts with parasites which are more specialized. Closeness centrality scores each node based on its relative distance to all other nodes in the network. It identifies hosts or parasites that are close to many other nodes. Nodes with high closeness scores are well-connected in the network and can act as “broadcasters” that are best positioned to influence the entire network. Proportional

generality is a measure of the number of partner species in relation to the number of potential partners. Species that utilize a large proportion of available partners (i.e., hosts or parasite species) have higher values of this network property. To incorporate host network indices into our parasite species-level database, each parasite was assigned the mean values of network centrality measures across all its hosts. For example, to assign a value for the “host closeness” variable of given parasite, we extracted and averaged the closeness scores of all its associated hosts. This resulted in six network variables for each parasite: three direct measures of parasite node centrality, and three measures of host node centrality aggregated to the parasite level, by taking the mean across all hosts of that parasite.

Host-parasite community analysis

Using the network structure described above, we defined two types of communities for each parasite species. Firstly, a parasite’s “host community” is its host range (i.e., all associated host species). Secondly, a parasite’s “parasite community” consists of the other parasites associated with host species in its host range. This is illustrated in Figure 2.2. For each parasite in the network, we calculated the parasite community size (number of species), proportion of parasites in the parasite community that were known to be zoonotic, and proportion of host species in the host community that are known carriers of zoonotic parasites. All zooscored parasites (protozoa and non-protozoa) and their hosts were considered when calculating the network metrics. We then filtered the data to include only network metrics protozoa species in our database.

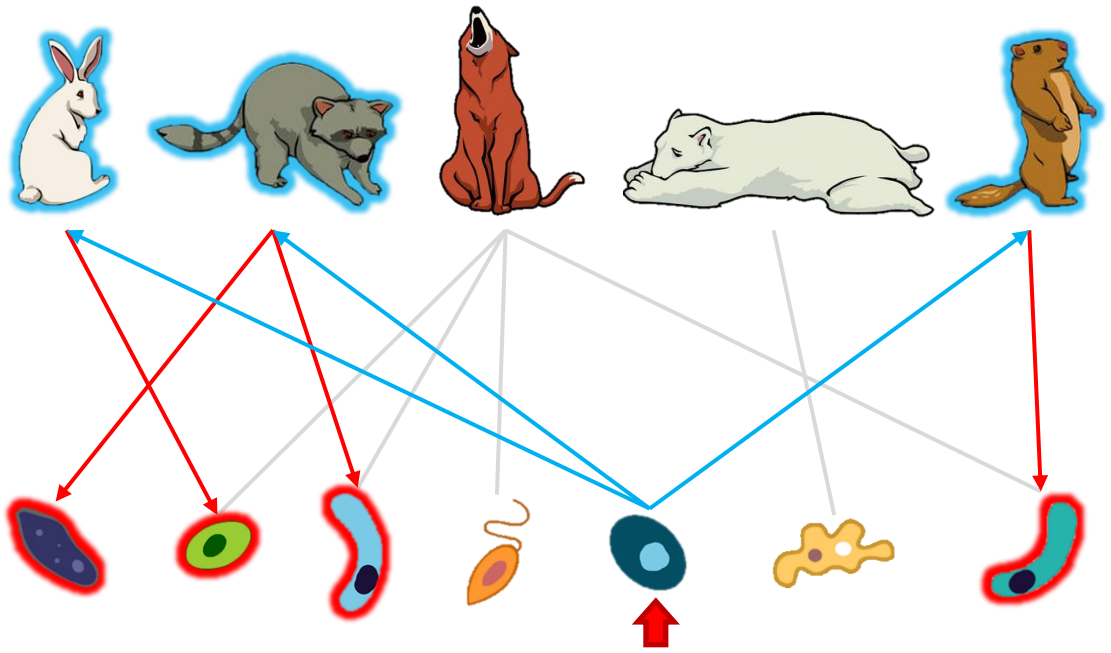


Figure 2.2. A host-parasite bipartite network in which host and parasite species are nodes of two separate modes. Edges represent a host-parasite association recorded in the GMPD. In this example, the focal parasite species (marked by a red arrow), is associated with three host species that make up its host "host community" (highlighted in blue). All other parasites associated with those hosts comprise the focal species' "parasite community" (highlighted in red). Note that a focal species is not included in its own parasite community.

Variable selection

We performed correlation analyses using the R package *corrplot* (Wei & Simko 2017) to create a correlation matrix and calculate Pearson's Correlation Coefficient (PCC) between all numeric variables in our dataset. Our original dataset included 48 variables: 14 intrinsic, 14 host-related, 7 environmental, 12 network and community traits, and 1 measure of research effort (WOS citation count). We removed variables that were highly correlated with others (PCC > 0.7 or < -0.7). We also excluded variables that were biologically redundant. For example, we chose to remove two out of the four parasite taxonomic levels, retaining phylum and order while excluding class and family.

Variables were also selected based on how complete the data were. We calculated coverage for all variables in our dataset and dropped those below 40%. Variables that remained after filtering by correlation and coverage were included as predictors in the model. The final dataset included a total 29 variables: 8 intrinsic traits, 12 host-related traits, 2 environmental traits, 6 network and community traits, and 1 measure of study effort. These variables were later regrouped into six categories. Information on these 29 variables, including definitions and percent coverage, is provided in Appendix A, Table A1.

Statistical modeling

Binary classification using boosted regression trees

Binary classification (or two-class classification) was performed using boosted regression trees (BRT), an ensemble learning method that can accommodate missing data, different classes of predictor variables, and non-linear relationships between

predictor and response variables (Elith *et al.* 2008). BRT analysis was performed using the machine learning algorithm XGBoost (eXtreme Gradient Boosting) with the *xgboost* packing in R (Chen *et al.* 2019). XGBoost was chosen because it optimizes standard gradient boosting machine algorithms by allowing for parallel processing to reduce computing time, built-in cross validation, efficient handling of missing values, tree-pruning, and regularization parameters to reduce overfitting.

Model fitting and parameter tuning

We divided our data into two subsets: a training set used to fit parameters and train the model, and a testing set into evaluate the model. We used the *caret* package (Kuhn 2008) to create a 65%/35% train-test split. The dataset was randomly sampled while preserving the proportion of positive and negative observations between training and testing sets. Because of the low ratio of zoonotic to non-zoonotic protozoa species in our dataset, we used tuning metrics best suited for classification of imbalanced data (Sun *et al.* 2009; Branco *et al.* 2015)

After partitioning the data into training and testing sets, we fit models to the training data with 5-fold cross-validation and a parameter grid search to determine the best parameter values for our model. Model parameters were selected to maximize the F_1 score, which is particularly useful for evaluation of imbalanced classification problems (i.e., unequal numbers of positive and negative classes) like ours, because it balances precision and recall (He & Ma 2013).

$$F_1 = 2 \times \frac{presicion \times recall}{precision + recall}$$

Precision is calculated as the number of accurately predicted positives divided by the number of total predicted positives. Recall is calculated as the number of accurately predicted positives divided by the number of actual positives in the dataset (also referred to as sensitivity or the true positive rate).

We then fit a final model using the optimal parameter values identified from the parameter grid search. All parameter values used in the model are reported in Table A2.

Model evaluation and predictions

We evaluated the performance of our model on the training and testing dataset. The BRT model predictions are a range of continuous values from 0 – 1, representing the probability that a protozoa species is in the zoonotic class. To convert these continuous prediction scores into binary classifications, we used a threshold that maximized the geometric mean or G-mean of our training data. G-Mean is a metric for imbalanced classification that, if optimized, seeks a balance between the sensitivity and the specificity (Ferri *et al.* 2009).

$$G - mean = \sqrt[2]{sensitivity \times specificity}$$

We then evaluated performance using the area under the receiver operating characteristic curve (AUC) on the raw scores and the F_1 -measure and accuracy on the thresholded, binary predictions. Accuracy was calculated as the total number of correct predictions divided by the total number of predictions.

Variable importance

Importance scores were calculated for each predictor variable used in the model, allowing them to be ranked and compared to each other. The scores indicate the value

of each variable for construction of boosted regression trees within the model; the more frequently a variable is used to make these decisions, the higher its relative importance. For an individual tree, a variable's importance was calculated based on how much it improved our model performance measure (AUC) at each split. Then the variable importance scores were averaged across all trees in the model to yield the overall variable importance scores for each predictor. This was calculated using the *xgb.importance* function in the *xgboost* package.

To assess the relative importance of the trait variables and study effort, we grouped them into the following five categories: parasite generalism, community traits, intrinsic parasite traits, host-related traits, environmental traits, and research effort. We permuted each category 128 times and estimated relative importance by calculating the decrease in AUC when each category was permuted, compared to the AUC of the model with no permutation. Partial dependance plots were used to visualize the marginal effect of the six most important variables. All variable importance calculations were done using the training dataset.

Results

Dataset summary

The 19 zoonotic protozoa species in our dataset were *Trypanosoma brucei, Toxoplasma gondii, Plasmodium brasilianum, Trypanosoma cruzi, Entamoeba coli, Balantidium coli, Trypanosoma rangeli, Leishmania shawi, Entamoeba chattoni, Plasmodium inui, Plasmodium cynomolgi, Plasmodium knowlesi, Sarcocystis hominis, Cryptosporidium parvum, Entamoeba hartmanni, Entamoeba polecki, Leishmania chagasi, Leishmania infantum,* and *Babesia EU2*.

Model evaluation and predictions

Our G-means approach to thresholding resulted in a threshold of 0.2048. We applied this threshold to both the training and testing data to calculate evaluation metrics that require binary predictions (F_1 measure and accuracy). Our model performed well on both the training data and testing data, with slightly lower performance on testing data (Table 2.1).

Table 2.1. Predictive performance of the BRT model on training and testing data evaluated by AUC, F_1, and accuracy.

Evaluation metric	Training data (n = 148)	Testing data (n = 78)
AUC	0.985	0.824
F1 measure	0.703	0.400
Accuracy	0.926	0.846

Our model was able to discriminate between zoonotic and non-zoonotic protozoa when applied to the full dataset with an accuracy of 85% (Table 2.1, Fig. 2.3). The five

protozoa most highly ranked by our model are all known to be zoonotic, and the majority of non-zoonotic protozoa were assigned low probabilities. Several notable protozoa species were ranked highly by our model (in order of highest probability): *Neospora caninum*, *Entamoeba histolytica, Entamoeba dispar, Leishmania donovani,* and *Leishmania braziliensis*. Some of these species are recognized in the literature as causing disease only rarely in humans, or human disease attributed to transmission from contaminated sources rather than from animal reservoirs. Zoonotic protozoa with a low predicted probability of directional spillover from animal hosts included (in order of lowest probability) *Babesia EU2, Plasmodium knowlesi, Sarcocystis hominis,* and *Trypanosoma rangeli*.

Importance of trait variables

Our final model included 11 of the original 29 variables we input into the model, including at least one variable from each trait category (Fig. 2.4). Probability of zoonotic status tended to increase with parasite generalism (Fig. 2.5c-e). Variables from the generalism category made up five of these 11 variables, and three of the top five variables (parasite betweenness, number of mammalian organ systems infected, and the number of transmission modes). This suggests the parasite generalism category is particularly important for predicting zoonotic status. Probability of zoonotic status also increased with increasing research effort Fig. 2.5b).

Order *Amoebida* was the strongest predictor variable (Fig 2.4), and protozoa species in this order had a higher probability of being zoonotic than those in other orders (Fig. 2.5a). There were six *Amoebida* species in our database, all in the genus *Entamoeba*. Four out of those six *Entamoeba* species are known to be zoonotic. The

other two, which are not known to be zoonotic, had high model predicted probabilities of being zoonotic (Fig 2.3). In fact, all *Entamoeba* species in our dataset had predicted probabilities over 0.5. In total, only 12 species ranked above 0.5.

The results of our permutation analysis (Fig. 2.6) mirrored those of our individual variable importance calculations. All categories of variables contributed somewhat to model performance, as illustrated by the reduction in performance following the permutation of each category (Fig. 2.6). Permuting variables in the generalism category resulted in the largest reduction in AUC, indicating this category is most important. This was followed by the sampling category, suggesting that research effort differs significantly between zoonotic protozoan and non-zoonotic protozoa. Permuting the hosts and community categories resulted in the lowest reductions in performance relative to the other categories. Intrinsic protozoan parasite traits and environmental traits had an intermediate effect on performance. They were significantly less influential than parasite generalism but more influential than host or community traits.

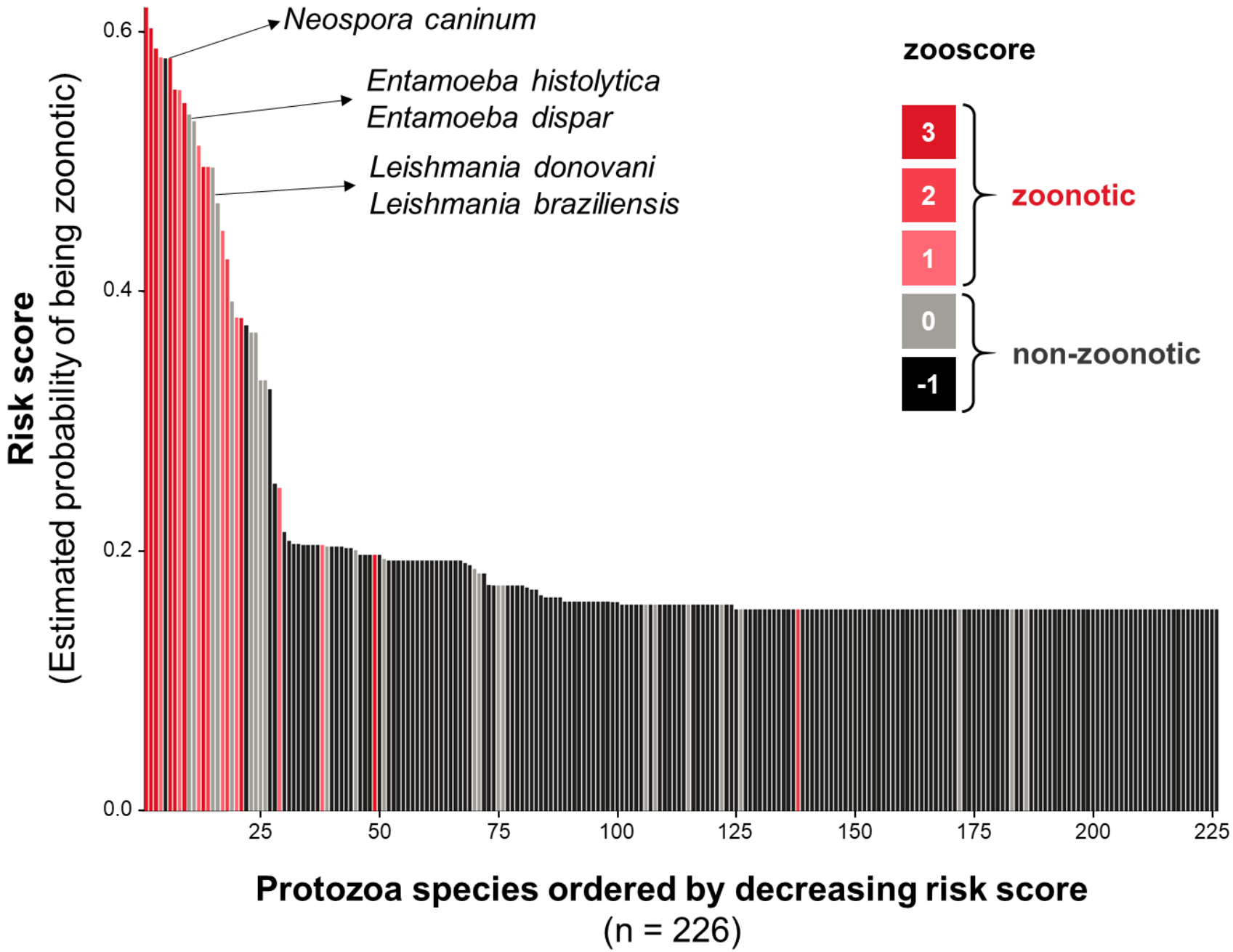

Neospora caninum
Entamoeba histolytica
Entamoeba dispar
Leishmania donovani
Leishmania braziliensis
zooscore
3
2
1
zoonotic
0
-1
non-zoonotic
Risk score
(Estimated probability of being zoonotic)
0.6
0.4
0.2
0.0
25
50
75
100
125
150
175
200
225
Protozoa species ordered by decreasing risk score
(n = 226)

Figure 2.3. Rank-ordered plot of average model-estimated probability of being zoonotic for each protozoa species. Black and grey bars represent species not known to be zoonotic and red bars represent species known to be zoonotic. Shades of color within each of the two categories indicate the zoonotic status code, or zooscore, assigned to each parasite species. Zooscores are defined in flowchart above. Parasites with zooscores of 1, 2, or 3 were considered zoonotic, and those with a zooscore of -1 or 0 were considered non-zoonotic. Bars corresponding to the five highest-ranking non-zoonotic protozoa species are labeled.

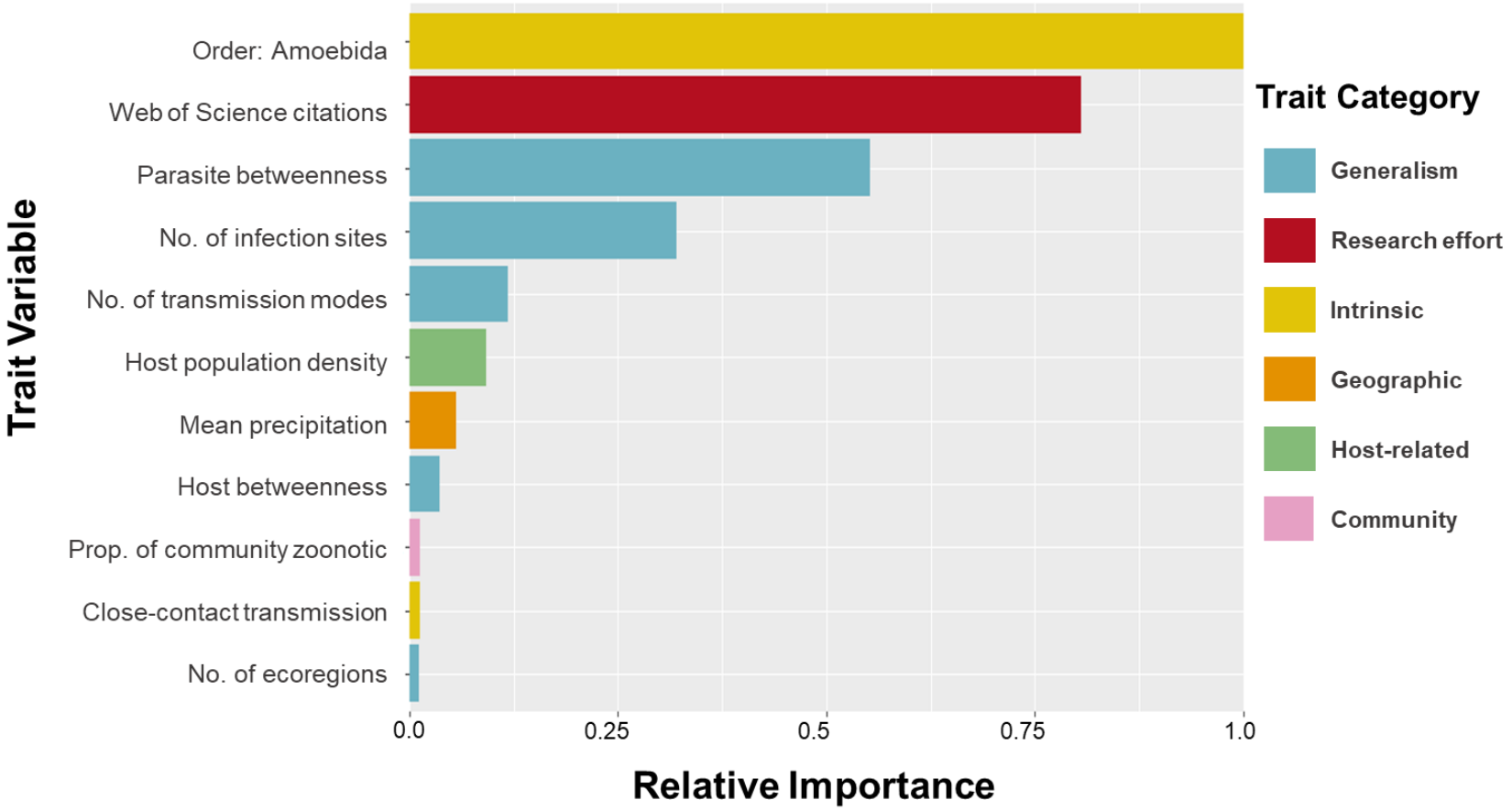


Figure 2.4. The relative importance of trait variables included in the BRT model for predicting the zoonotic status of protozoan parasites. Variables are ordered based on the decreasing importance scores, and x-axis values correspond to the relative improvement in model performance as measured by AUC, averaged across all trees and scaled to the most important variable. The bars are colored by trait category.

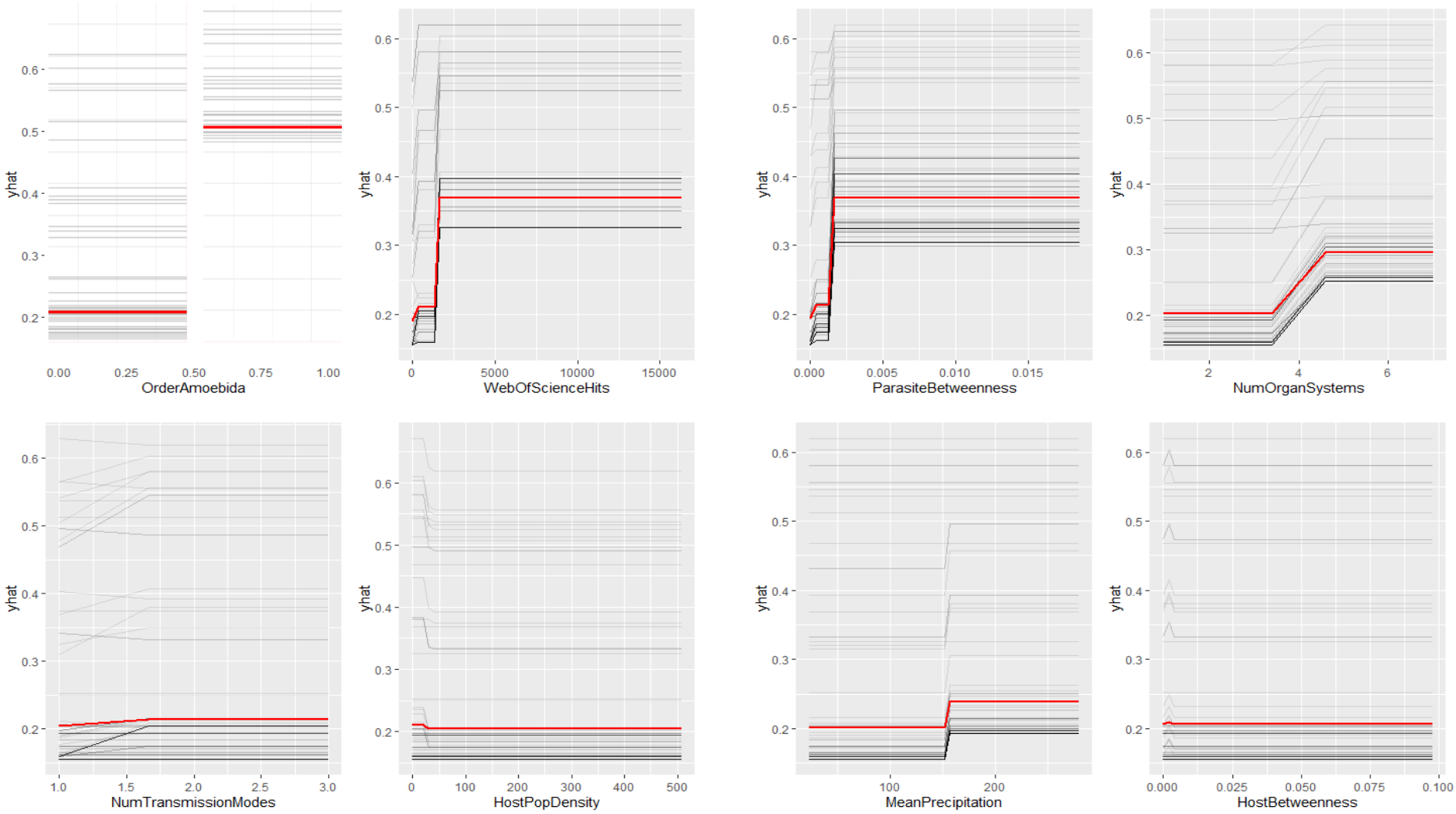


Figure 2.5. Partial dependence plots of the top 8 predictor variables. Plots appear in order of predictive importance, from left to right and top to bottom. The grey lines represent the average marginal effect of each variable on the predicted

probability of being zoonotic for each protozoa species in the dataset, and the red lines correspond to the average across all 226 species.

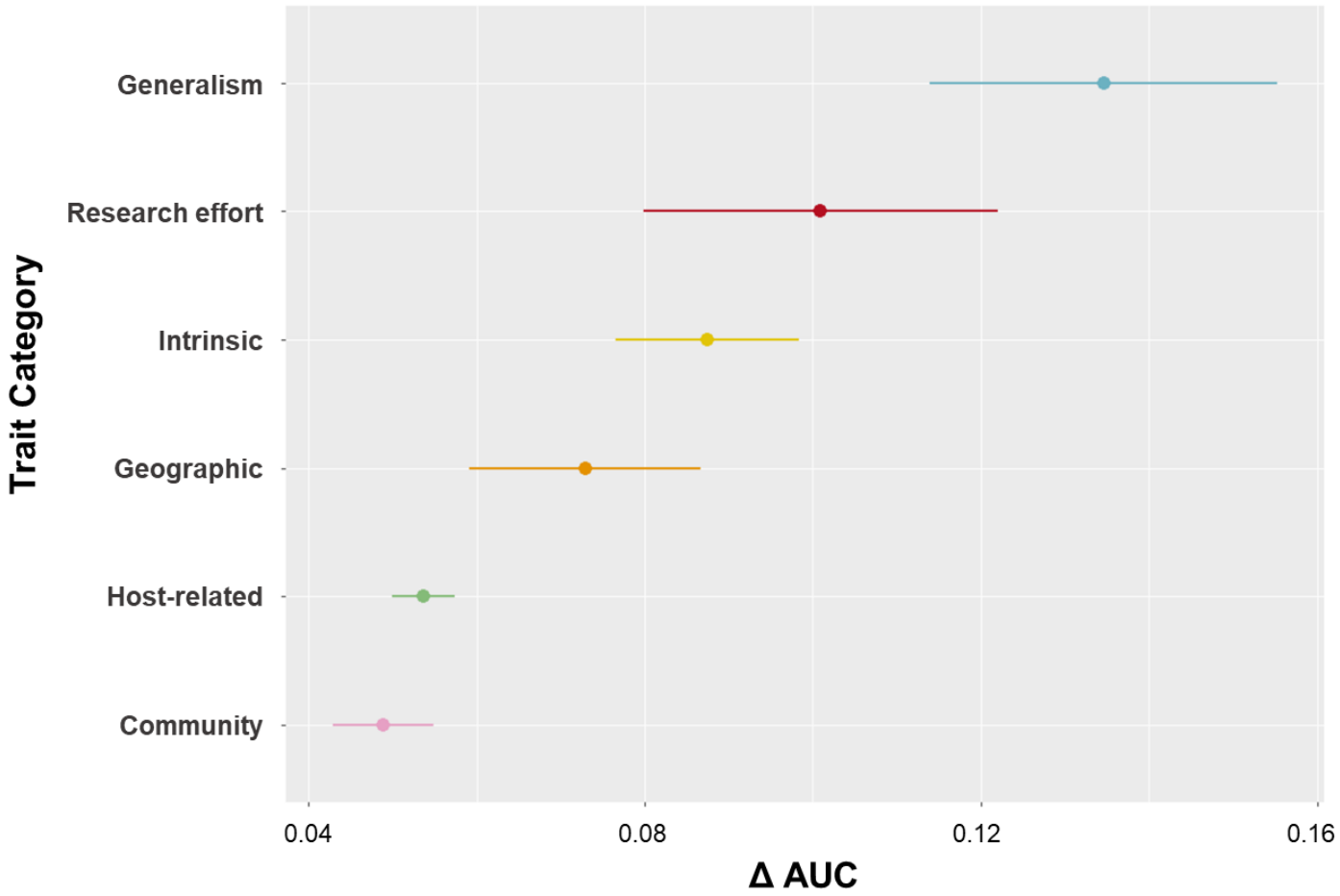


Figure 2.6. The influence of each category of trait variables on model predictive performance, as measured by the decrease in AUC caused by permuting all variables in that category, relative to the AUC of the full model. Larger values of Δ AUC indicate that that the variables in that category are more important for correctly distinguishing zoonotic and non-zoonotic protozoa. Points represent the mean and error bars represent the standard deviation across 128 permutations.

Discussion

In this study, we show that the zoonotic risk of protozoan parasites can be predicted with a high degree of accuracy using a statistical model trained on known traits of parasites, hosts, and host-parasite interaction networks. We identify that parasite generalism and taxonomy are the most important characteristics that distinguish zoonotic from non-zoonotic protozoa, and we use this information to identify protozoa species that are high-risk candidates for undiscovered current or future zoonoses. We suggest that these species should be priority targets for future surveillance.

Predictions

Our model was able to accurately distinguish between protozoa species known to be zoonotic and those not known to be zoonotic (Table 2.1). This is illustrated by the high average risk score of known zoonotic parasites and average low risk score of parasites not known to be zoonotic (Fig. 2.3; Appendix B, Fig. A1). The three species with the highest zoonotic risk scores were: *Entamoeba coli* (0.62), *Trypanosoma brucei* (0.60), and *Toxoplasma gondii* (0.59). Not only are these protozoa all zoonotic, but they also all have a zooscore of 3 (i.e., highest zoonotic potential), further evidence of the robustness of our model fit.

However, some protozoa species that are known to be zoonotic had notably low risk scores: *Babesia EU2*, *Plasmodium knowlesi*, and *Sarcocystis hominis* (Fig. 2.3; Appendix B, Table A5). A possible explanation for this is that we had relatively little data for these protozoa. *B. EU2* and *S. hominis* each appeared once in the GMPD, and *P. knowlesi* appeared 10 times and had only three associated hosts. One important

predictor of parasite zoonotic status was betweenness centrality, but betweenness is sensitive to sampling effort, because a parasite that has been studied more often will have more of its true associations represented in the host-parasite bipartite network. Therefore, zoonotic parasites that were relatively less studied in wildlife – and therefore not well represented in the GMPD – likely had artificially low betweenness scores and thus lower predicted zoonotic potential. Indeed, *Babesia EU2* and *Sarcosystis hominis* had betweenness centralities of zero in our dataset. Because sampling effort had high importance in our model, the anomalously low estimated probabilities for these species is likely due to the bias for more well-studied species to have more documented host-parasite associations, and also to be zoonotic and thus also have higher estimated probabilities. The performance of this model could be further improved with more complete data for poorly sampled protozoan parasites.

Our model also identified a number of non-zoonotic protozoa (zooscore < 1) with high estimated probabilities, suggesting high zoonotic potential for these parasites. The protozoa species currently not known to be zoonotic that had notably high risk scores were: *Neospora caninum*, *Entamoeba histolytica, Entamoeba dispar, Leishmania donovani,* and *Leishmania braziliensis* (Fig. 2.3, Appendix B, Table A5). *N. caninum* had a zooscore of -1 (observed in animals only), while the rest had zooscores of 0 (observed in both animals and humans, but no confirmed zoonotic transmission). That many of the high-ranking “non-zoonotic” protozoa are found in both animals and humans suggests that they might pose a higher threat of spilling over from wild animals into human populations because they are already adapted to infecting human tissues.

Conversely, it is possible that these are cases of reverse zoonosis or “spill back” transmission from humans to wild animals.

The parasite with the highest estimated probability that is not known to currently infect humans (zooscore of -1) was *Neospora caninum.* The definitive hosts of this parasite are wild canids and domestics dogs (Rosypal & Lindsay 2005; Gondim 2006; King *et al.* 2011; Almería 2013), and it is transmitted within and across species via a fecal-oral route. Until 1988, *N. caninum* was misdiagnosed as *Toxoplasma gondii* in animals, because of similarities in ultrastructure, genetics, and pathology (Dubey & Lindsay 1996; Tranas *et al.* 1999; Almería 2013). It could be that a similar misdiagnoses are happening in humans, meaning that the parasite is already present in human populations but has not yet been detected. Of the two past studies that looked for serological evidence of human exposure, one found evidence for exposure and the other found none (Tranas *et al.* 1999; McCann *et al.* 2008); both suggest further study and vigilance to the possibility of human infection, as do the results from our model.

Two other high-ranking parasite species in the model were *E. histolytica* and *E. dispar,* both of which have a zooscore of 0. These parasites are closely related (Dong *et al.* 2017) and in the order *Amoebida*. Despite the high risk score assigned to *E. histolytica*, there is little to no evidence for zoonotic infection with *E. histolytica* in humans (Dubey 2003; Mak 2004; Junaidi *et al.* 2020). Past reports of zoonotic infection were later revealed to be misdiagnoses of *E. chattoni* (Sargeaunt *et al.* 1992). Similarly, to our knowledge there is no evidence of zoonotic transmission for *E. dispar* (Mak 2004). For both parasites, zoonotic transmission could be limited because the parasite rarely encysts in the lumen of animals, and cyst formation is required for onward

transmission (Junaidi *et al.* 2020). Taken together, these data suggest that there is uncertainty about the zoonotic potential of these two parasites. Limited serological evidence points to human exposure, and high trait similarity to known zoonotic pathogen suggests that zoonotic transmission from wildlife to humans via domestic animals is not implausible, but to date there is no evidence that they are zoonotic. We suggest that these species should be targets of continued surveillance.

Of the non-zoonotic protozoa, most were found to infect both animals and humans (zooscore = 0) and very few were animal-only protozoa (zooscore = -1). This suggests that protozoa that are capable of infecting both humans and animals are more likely to spill over, but those that only infect animals unlikely to become zoonotic in the future. One explanation for this might be that protozoa have low adaptability compared to other parasite taxa such as virus or bacteria, which exhibit high mutation rates and horizontal gene transfer (Woolhouse *et al.* 2005). Due to the relative genetic "stability" of protozoa, it is possible that animal-only protozoa are unlikely to adapt to a new human host to cross the species barrier into humans.

Importance of trait variables

The high relative importance of the order *Amoebida* in predicting the zoonotic status of protozoa species could have to do with the distribution of parasites in orders in the dataset. *Amoebida* is the order with the highest proportion of zoonotic parasites in the dataset (beyond those with just one parasite in the order).

Variables in the generalism category, particularly parasite betweenness centrality and number of organ systems the parasite is known to infect, were the most important for predicting the zoonotic status of protozoan parasites (Fig. 2.4; Fig. 2.6). Compared

to other parasite taxa, protozoa infect a smaller phylogenetic range of hosts (Park *et al.* 2018). Zoonotic protozoa may be an exception to this trend because they have a higher degree of generalism compared to protozoa on average.

The most important generalist trait was network betweenness of the protozoa. Interestingly, other calculated network metrics, closeness and proportional generality, contributed less to model performance. This suggests that what is important is not the connectedness of the focal parasite itself, but rather the role it plays in connecting hosts with isolated, distinct parasite ranges, acting as a "bridge". The role of species acting as a "bridge" for disease transmission has been proposed for hosts, particularly for domestic mammals serving as bridges between wildlife and humans (Caron *et al.* 2015; Berrian *et al.* 2016). Although the mechanism for "bridge parasites" is clearly different, a species' ability to infect otherwise isolated hosts indicates a high plasticity in host preference, and therefore a higher zoonotic risk.

Limitations

This study could be improved by obtaining host-parasite association records that are not limited to the host clades in the GMPD. Though the GMPD provides valuable data on the parasites of wild primates, ungulates and carnivores, we are missing other mammalian host clades – such as rodents, bats, and domestic animals – which can be important reservoirs of zoonotic pathogens (Calisher *et al.* 2006; Luis *et al.* 2013; Plourde *et al.* 2017). In addition to mammals, birds, reptiles, and other invertebrate hosts are known to harbor zoonoses (Esch & Petersen 2013; Viana *et al.* 2014). Broadening the range of hosts in this study would add new host-parasite associations, allowing us to more accurately estimate network centrality and other measures of

generalism. For example, *Sarcocystis hominis* has a broad host range, of which many are domestic animals not included in our study (Tenter 1995). Such parasites will have artificially low measures of generalism in our dataset.

Another limitation is we had incomplete coverage of some covariates (Appendix A, Table A1). This is important to note, especially given the importance of the network metrics, which are built on networks that we know have missing data.

Conclusions

The zoonotic risk of protozoan parasites was predicted with a high degree of accuracy using a statistical model trained on known traits of parasites, hosts, and host-parasite interaction networks. The most important traits for distinguishing zoonotic protozoa from non-zoonotic protozoa were measures of parasite generalism, taxonomic order, and research effort. This information was used to identify protozoa species that are high-risk candidates for undiscovered current or future zoonoses. Non-zoonotic protozoa with high zoonotic potential were closely related to known zoonotic parasites, or found to infect both humans and wildlife. We suggest that these species should be priority targets for future surveillance.

CHAPTER 3

CONCLUSIONS

Understanding the traits of zoonotic protozoan parasites is important for predicting and controlling emerging infectious diseases. In this study, we examined the relative importance of different traits for predicting the zoonotic status of parasitic protozoa, and found that variables in the generalism category, particularly parasite betweenness centrality and number of organ systems the parasite is known to infect, were the most important for predicting the zoonotic status of protozoan parasites. This corroborated our hypothesis that protozoa species with a greater host range, geographic range, number of transmission modes, or sites of infection are more prone to cross-species transmission.

We found that a statistical model trained on known traits of parasites, hosts, and host-parasite interaction networks can accurately estimate the zoonotic risk of protozoan parasites. Our model was able to accurately distinguish between protozoa species known to be zoonotic and those not known to be zoonotic. Our model also identified a number of protozoa species which are currently not known to be zoonotic but had high estimated probabilities of being zoonotic. Many of these are closely related to known zoonotic protozoa. Additionally, most are already known to infect both animals and humans, though direct spillover has never been documented. We suggest that these protozoa species of wild mammals with high model-estimated zoonotic potential should be targets of surveillance efforts.

REFERENCES

Acha, P.N., Szyfres, B. & Acha, P.N. (2003). *Parasitoses*. Zoonoses and communicable diseases common to man and animals / [Pedro N. Acha; Boris Szyfres]. 3. ed. Pan American Health Organization, Washington, DC.

Ahmed, M.A. & Cox-Singh, J. (2015). *Plasmodium knowlesi* – an emerging pathogen. *ISBT Sci Ser*, 10, 134–140.

Allen, T., Murray, K.A., Zambrana-Torrelio, C., Morse, S.S., Rondinini, C., Marco, M.D., *et al.* (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8, 1124.

Almería, S. (2013). *Neospora caninum* and Wildlife. *ISRN Parasitol*, 2013.

Altizer, S., Ostfeld, R.S., Johnson, P.T.J., Kutz, S. & Harvell, C.D. (2013). climate change and infectious diseases: from evidence to a predictive framework. *Science*, 341, 514–519.

Bender, J.B., Hueston, W. & Osterholm, M. (2006). Recent animal disease outbreaks and their impact on human populations. *Journal of Agromedicine*, 11, 5–15.

Berrian, A.M., van Rooyen, J., Martínez-López, B., Knobel, D., Simpson, G.J.G., Wilkes, M.S., *et al.* (2016). One Health profile of a community at the wildlife-domestic animal interface, Mpumalanga, South Africa. *Preventive Veterinary Medicine*, 130.

Branco, P., Torgo, L. & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions.

Brock, P.M., Fornace, K.M., Parmiter, M., Cox, J., Drakeley, C.J., Ferguson, H.M., *et al.* (2016). Plasmodium knowlesi transmission: integrating quantitative approaches from epidemiology and ecology to understand malaria as a zoonosis. *Parasitology*, 143, 389–400.

Calisher, C.H., Childs, J.E., Field, H.E., Holmes, K.V. & Schountz, T. (2006). Bats: Important Reservoir Hosts of Emerging Viruses. *Clinical Microbiology Reviews*, 19, 531–545.

Caron, A., Cappelle, J., Cumming, G.S., de Garine-Wichatitsky, M. & Gaidet, N. (2015). Bridge hosts, a missing link for disease ecology in multi-host systems. *Veterinary Research*, 46, 83.

Chen, T., He, T. & Khotilovich, V. (2019). *xgboost: Extreme Gradient Boosting*.

Cleaveland, S., Laurenson, M.K. & Taylor, L.H. (2001). 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*, 356, 991–999.

Cox-Singh, J., Davis, T.M.E., Lee, K.-S., Shamsul, S.S.G., Matusop, A., Ratnam, S., *et al.* (2008). *Plasmodium knowlesi* Malaria in humans is widely distributed and potentially life threatening. *Clinical Infectious Diseases*, 46, 165–171.

Craft, M.E. (2015). Infectious disease transmission and contact networks in wildlife and livestock. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370.

Dallas, T.A., Han, B.A., Nunn, C.L., Park, A.W., Stephens, P.R. & Drake, J.M. (2019). Host traits associated with species roles in parasite sharing networks. *Oikos*, 128, 23–32.

Dong, H., Li, J., Qi, M., Wang, R., Yu, F., Jian, F., *et al.* (2017). Prevalence, molecular epidemiology, and zoonotic potential of Entamoeba spp. in nonhuman primates in China. *Infection, Genetics and Evolution*, 54, 216–220.

Dormann, C., Gruber, B. & Fründ, J. (2008). Introducing the bipartite Package: Analysing Ecological Networks. *R News*.

Dubey, J.P. (2003). Review of *Neospora caninum* and neosporosis in animals. *Korean J Parasitol*, 41, 1–16.

Dubey, J.P. & Lindsay, D.S. (1996). A review of *Neospora caninum* and neosporosis. *Vet Parasitol*, 67, 1–59.
Elith, J., Leathwick, J.R. & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802–813.
Esch, K.J. & Petersen, C.A. (2013). Transmission and Epidemiology of Zoonotic Protozoal Diseases of Companion Animals. *Clinical Microbiology Reviews*, 26, 58–85.
Evans, M.V., Murdock, C.C. & Drake, J.M. (2018). Anticipating emerging mosquito-borne flaviviruses in the USA: What comes after Zika? *Trends in Parasitology*, 34, 544–547.
Faurby, S., Davis, M., Pedersen, R.Ø., Schowanek, S.D., Antonelli1, A. & Svenning, J.-C. (2018). PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology. *Ecology*, 99, 2626–2626.
Feng, Y. & Xiao, L. (2011). Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis. *Clin Microbiol Rev*, 24, 110–140.
Ferri, C., Hernández-Orallo, J. & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30, 27–38.
Flanagan, M.L., Parrish, C.R., Cobey, S., Glass, G.E., Bush, R.M. & Leighton, T.J. (2012). Anticipating the species jump: Surveillance for emerging viral threats. *Zoonoses and Public Health*, 59, 155–163.
Fong, I.W. (2017). Animals and mechanisms of disease transmission. In: *Emerging Zoonoses: A Worldwide Perspective*, Emerging Infectious Diseases of the 21st Century (ed. Fong, I.W.). Springer International Publishing, Cham, pp. 15–38.
Fornace, K.M., Brock, P.M., Abidin, T.R., Grignard, L., Herman, L.S., Chua, T.H., *et al.* (2019). Environmental risk factors and exposure to the zoonotic malaria parasite *Plasmodium knowlesi* across northern Sabah, Malaysia: a population-based cross-sectional survey. *The Lancet Planetary Health*, 3, e179–e186.
Gondim, L.F.P. (2006). *Neospora caninum* in wildlife. *Trends in Parasitology*, 22, 247–252.
Gortazar, C., Reperant, L.A., Kuiken, T., de la Fuente, J., Boadella, M., Martínez-Lopez, B., *et al.* (2014). Crossing the interspecies barrier: opening the door to zoonotic pathogens. *PLoS Pathogens*, 10, e1004129.
Han, B.A., Majumdar, S., Calmon, F.P., Glicksberg, B.S., Horesh, R., Kumar, A., *et al.* (2019). Confronting data sparsity to identify potential sources of Zika virus spillover infection among primates. *Epidemics*, 27, 59–65.
Han, B.A., O'Regan, S.M., Schmidt, J.P. & Drake, J.M. (2020). Integrating data mining and transmission theory in the ecology of infectious diseases. *Ecology Letters*, 23, 1178–1188.
Han, B.A., Schmidt, J.P., Bowden, S.E. & Drake, J.M. (2015). Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences*, 112.
He, H. & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons.
Herricks, J.R., Hotez, P.J., Wanga, V., Coffeng, L.E., Haagsma, J.A., Basáñez, M.-G., *et al.* (2017). The global burden of disease study 2013: What does it mean for the NTDs? *PLOS Neglected Tropical Diseases*, 11, e0005424.
Huber, C., Finelli, L. & Stevens, W. (2018). The economic and social burden of the 2014 Ebola outbreak in West Africa. *The Journal of Infectious Diseases*, 218, S698–S704.
Institute for Health Metrics and Evaluation (IHME). (2019). Global Burden of Disease Study 2019 (GBD 2019).
Isaac, N.J.B., Turvey, S.T., Collen, B., Waterman, C. & Baillie, J.E.M. (2007). Mammals on the EDGE: Conservation of priorities based on threat and phylogeny. *PLOS ONE*, 2, e296.
Johnson, C.K., Hitchens, P.L., Evans, T.S., Goldstein, T., Thomas, K., Clements, A., *et al.* (2015). Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific Reports*, 5, srep14830.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L., *et al.* (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90, 2648–2648.
Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., *et al.* (2008). Global trends in emerging infectious diseases. *Nature*, 451, 990–993.
Junaidi, J., Cahyaningsih, U., Purnawarman, T., Latif, H., Sudarnika, E., Hayati, Z., *et al.* (2020). *Entamoeba histolytica* neglected tropical diseases (NTD) agents that infect humans and some other mammals: A review. *E3S Web Conf.*, 151, 01019.
Karesh, W.B., Dobson, A., Lloyd-Smith, J.O., Lubroth, J., Dixon, M.A., Bennett, M., *et al.* (2012). Ecology of zoonoses: natural and unnatural histories. *The Lancet*, 380, 1936–1945.
King, J.S., Jenkins, D.J., Ellis, J.T., Fleming, P., Windsor, P.A. & Šlapeta, J. (2011). Implications of wild dog ecology on the sylvatic and domestic life cycle of *Neospora caninum* in Australia. *The Veterinary Journal*, 188, 24–33.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1–26.
Leggett, H.C., Buckling, A., Long, G.H. & Boots, M. (2013). Generalism and the evolution of parasite virulence. *Trends in Ecology & Evolution*, 28, 592–596.
Luis, A.D., Hayman, D.T.S., O'Shea, T.J., Cryan, P.M., Gilbert, A.T., Pulliam, J.R.C., *et al.* (2013). A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proceedings of the Royal Society B: Biological Sciences*, 280, 20122753.
Luis, A.D., O'Shea, T.J., Hayman, D.T.S., Wood, J.L.N., Cunningham, A.A., Gilbert, A.T., *et al.* (2015). Network analysis of host-virus communities in bats and rodents reveals determinants of cross-species transmission. *Ecology Letters*, 18.
Mak, J.W. (2004). Important zoonotic intestinal protozoan parasites in Asia. *Trop Biomed*, 21, 39–50.
Martins, S.B., Häsler, B. & Rushton, J. (2015). Economic Aspects of Zoonoses: Impact of Zoonoses on the Food Industry. In: *Zoonoses - Infections Affecting Humans and Animals: Focus on Public Health Aspects* (ed. Sing, A.). Springer Netherlands, Dordrecht, pp. 1107–1126.
McCann, C.M., Vyse, A.J., Salmon, R.L., Thomas, D., Williams, D.J.L., McGarry, J.W., *et al.* (2008). Lack of serologic evidence of *Neospora caninum* in humans, England. *Emerg Infect Dis*, 14, 978–980.
Morse, S.S. (1995). Factors in the emergence of infectious diseases. *Emerg Infect Dis*, 1, 7–15.
Morse, S.S., Mazet, J.A., Woolhouse, M., Parrish, C.R., Carroll, D., Karesh, W.B., *et al.* (2012). Prediction and prevention of the next pandemic zoonosis. *The Lancet*, 380, 1956–1965.
Nii-Trebi, N.I. (2017). Emerging and Neglected Infectious Diseases: Insights, Advances, and Challenges. *BioMed Research International*, 2017, 5245021.
Nunn, C.L. & Altizer, S.M. (2005). The global mammal parasite database: An online resource for infectious disease records in wild primates. *Evolutionary Anthropology: Issues, News, and Reviews*, 14, 1–2.
Olival, K.J., Hosseini, P.R., Zambrana-Torrelio, C., Ross, N., Bogich, T.L. & Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature*, 546, 646–650.
Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., *et al.* (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51, 933–938.
Pandit, P. & Han, B.A. (2020). Rise of machines in disease ecology. *The Bulletin of the Ecological Society of America*, 101, e01625.
Park, A.W. (2019). Phylogenetic aggregation increases zoonotic potential of mammalian viruses. *Biology Letters*, 15, 20190668.

Park, A.W., Farrell, M.J., Schmidt, J.P., Huang, S., Dallas, T.A., Pappalardo, P., *et al.* (2018). Characterizing the phylogenetic specialism–generalism spectrum of mammal parasites. *Proceedings of the Royal Society B: Biological Sciences*, 285, 20172613.
Pisarski, K. (2019). The global burden of disease of zoonotic parasitic diseases: top 5 contenders for priority consideration. *Trop Med Infect Dis*, 4.
Plourde, B.T., Burgess, T.L., Eskew, E.A., Roth, T.M., Stephenson, N. & Foley, J.E. (2017). Are disease reservoirs special? Taxonomic and life history characteristics. *PLoS ONE*, 12.
Plowright, R.K., Reaser, J.K., Locke, H., Woodley, S.J., Patz, J.A., Becker, D.J., *et al.* (2021). Land use-induced spillover: a call to action to safeguard environmental, animal, and human health. *The Lancet Planetary Health*, 5, e237–e245.
Rosypal, A.C. & Lindsay, D.S. (2005). The sylvatic cycle of *Neospora caninum*: where do we go from here? *Trends in Parasitology*, 21, 439–440.
Ryan, U. & Cacciò, S.M. (2013). Zoonotic potential of *Giardia*. *International Journal for Parasitology*, Zoonoses Special Issue, 43, 943–956.
Salkeld, D.J., Stapp, P., Tripp, D.W., Gage, K.L., Lowell, J., Webb, C.T., *et al.* (2016). Ecological traits driving the outbreaks and emergence of zoonotic pathogens. *BioScience*, 66.
Sargeaunt, P.G., Patrick, S. & O'Keeffe, D. (1992). Human infections of *Entamoeba chattoni* masquerade as *Entamoeba histolytica*. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 86, 633–634.
Schmeller, D.S., Courchamp, F. & Killeen, G. (2020). Biodiversity loss, emerging pathogens and human health risks. *Biodivers Conserv*, 29, 3095–3102.
Singh, B., Sung, L.K., Matusop, A., Radhakrishnan, A., Shamsul, S.S., Cox-Singh, J., *et al.* (2004). A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *The Lancet*, 363, 1017–1024.
Stephens, P.R., Pappalardo, P., Huang, S., Byers, J.E., Farrell, M.J., Gehman, A., *et al.* (2017). Global Mammal Parasite Database version 2.0. *Ecology*, 98, 1476–1476.
Sun, Y., Wong, A.K.C. & Kamel, M.S. (2009). Classification of imbalanced data: a review. *Int. J. Patt. Recogn. Artif. Intell.*, 23, 687–719.
Taylor, L.H., Latham, S.M. & Woolhouse, M.E. (2001). Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci*, 356, 983–989.
Tenter, A.M. (1995). Current research on Sarcocystis species of domestic animals. *International Journal for Parasitology*, 25, 1311–1330.
Torgerson, P.R. & Macpherson, C.N.L. (2011). The socioeconomic burden of parasitic zoonoses: Global trends. *Veterinary Parasitology*, Special issue: Zoonoses in a Changing World, 182, 79–95.
Tranas, J., Heinzen, R.A., Weiss, L.M. & McAllister, M.M. (1999). Serological evidence of human infection with the protozoan *Neospora caninum*. *Clin Diagn Lab Immunol*, 6, 765–767.
Viana, M., Mancy, R., Biek, R., Cleaveland, S., Cross, P.C., Lloyd-Smith, J.O., *et al.* (2014). Assembling evidence for identifying reservoirs of infection. *Trends in Ecology & Evolution*, 29, 270–279.
Walker, J.W., Han, B.A., Ott, I.M. & Drake, J.M. (2018). Transmissibility of emerging viral zoonoses. *PLOS ONE*, 13, e0206926.
Wei, T. & Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix*.
Wolfe, N.D., Dunavan, C.P. & Diamond, J. (2007). Origins of major human infectious diseases. *Nature*, 447, 279–283.
Woolhouse, M.E.J., Haydon, D.T. & Antia, R. (2005). Emerging pathogens: the epidemiology and evolution of species jumps. *Trends in Ecology & Evolution*, 20, 238–244.

Yu, V.L. & Edberg, S.C. (2005). Global Infectious Diseases and Epidemiology Network (GIDEON): A world wide web-based program for diagnosis and informatics in infectious diseases. *Clinical Infectious Diseases*, 40, 123–126.

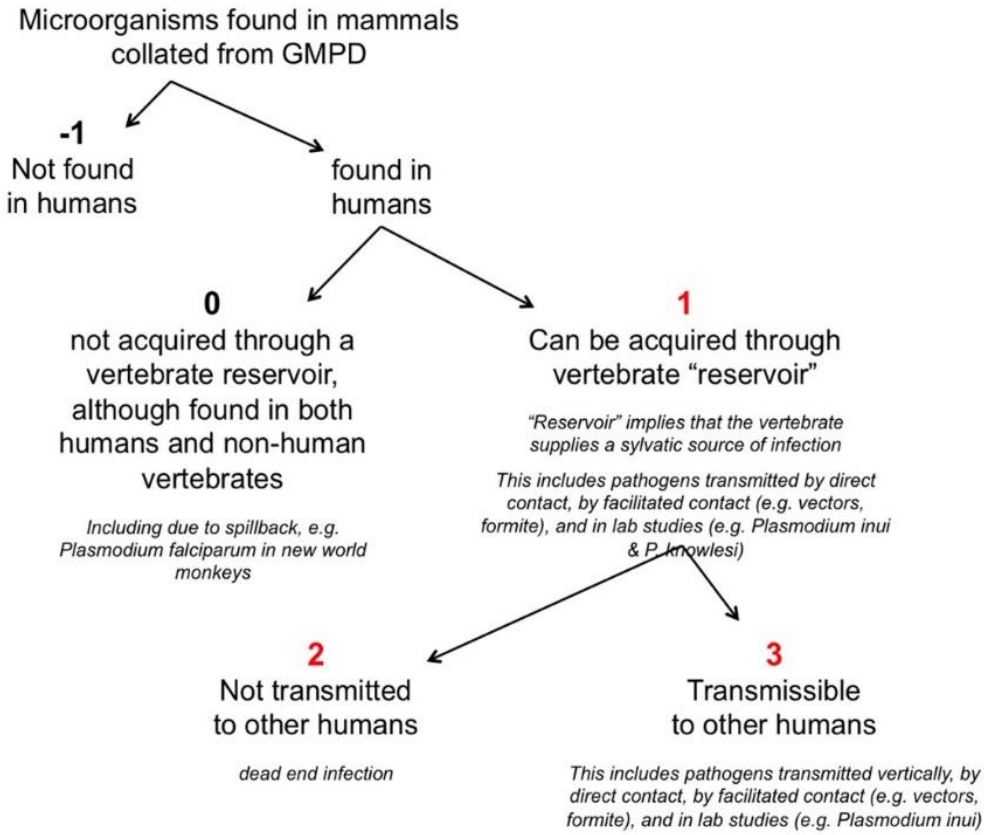
APPENDICES

APPENDIX A: METHODS

Protocol for assigning zoonotic status codes to GMPD parasites

Coding flowchart

Start at the top of the flowchart and follow the branches down as far as you can.



General instructions

1) To assess whether the microbe has been reported in humans, do the following:
   a) GIDEON database (Yu & Edberg 2005).
      i) if reported in the Diseases tab of GIDEON you know that it is pathogenic (move on to step 3).
      ii) if not reported in GIDEON, go to step 1b.
   b) Acha PAHO reference (Acha *et al.* 2003).
      i) if reported here, the microbe could be a pathogen or a commensal - read carefully to distinguish, and track down any primary references (step 4).

ii) If not reported in Acha PAHO, try 1c.

c) Search Google Scholar and/or Web of Science
   i) record the search string used and the number of documents checked (I searched using *"Acantheocephalus anguillae" human infection*; there were 48 papers returned in Google Scholar".
   ii) search through these references to find evidence that this parasite has been found in a human.
   iii) If you cannot find a reference for the occurrence of this microbe in a human, code it as -1 in the spreadsheet.
   iv) Record a confidence score for this parasite (see metadata tab in the spreadsheet).

2) Example: For *Acanthocephalus anguillae* I would record a zoonotic score of -1 (not found in humans), a confidence score of 3 (very little studied – data deficient).

3) If the microbe <u>has</u> been reported in humans, assess whether it is pathogenic or commensal by checking the following references (in order):
   a) If identified in GIDEON, the parasite is pathogenic.
   b) If found in the Acha PAHO reference, it could be either. Read carefully to distinguish.
   c) In all cases, track back to the *primary reference* reporting the parasite as pathogenic.

4) If the microbe is indeed a human pathogen, determine next whether there is evidence to suggest that the pathogen is transmitted from the animal to humans (code as 1 and record a confidence score).
   a) Check Acha PAHO for information to make this distinction.
   b) Check the primary literature (Google Scholar and/or Web of Science) and record SearchString and the number of hits in the spreadsheet.
   c) If there is no evidence to suggest that a vertebrate transmits infection to humans, code as 0 and record a confidence score.
5) If the parasite is known to be transmitted by a vertebrate reservoir, determine next whether there is evidence to suggest humans are passing infection on to other humans, or whether humans are only a dead-end host.
   a) Check Acha PAHO for information to make this distinction.
   b) Check the primary literature (Google Scholar and/or Web of Science)
   c) Assign a score of 2, 3, or 4 with confidence scores and citations in Zotero to justify the assigned scores.

6) If you assign a score based on what is reported in a particular paper and you would like me to double check it, flag the papers in Zotero by adding the tag "double check". In addition, write a short note in the Notes tab of Zotero about what specific uncertainties you had about the paper/parasite/score assignment.

7) Initial the parasite species you've just scored (WhoBy), enter the date of completion (DateEntry).

Table A1. Variable definitions and coverage

	Predictor	Type	Category	Description	Units	Data source	% Coverage (all)	% Coverage (zoonotic)
1	Number of Transmission Modes	intrinsic	generalism	Number of transmission modes the parasite is known to use. Categories are close-contact, non-close contact, intermediate, and vector transmission.	#	GMPD trait	77.0%	94.7%
2	Host Diet Breadth	host-related	generalism	Number of dietary categories eaten by each host species. Categories were defined as vertebrate, invertebrate, fruit, flowers/nectar/pollen, leaves/branches/bark, seeds, grass and roots/tubers.	#	PanTHERIA	96.9%	94.7%
3	Host Habitat Breadth	host-related	generalism	Number of habitat layers used by each host species. Categories were defined as above ground dwelling, aquatic, fossorial and ground dwelling.	#	PanTHERIA	81.9%	78.9%

4	Host Geographic Range	host-related	generalism	Host geographic ranges	km^2	Dallas et al. 2018	98.7%	100.0%
5	Host Island Endemicity	host-related	generalism	Classification scores correspond to four degrees of geographic isolation: occurs only on isolated islands, occurs on small land bridge islands, occurs on large land bridge islands, occurs on mainland	factor	PHYLACINE	98.7%	100.0%
6	Number of Organ Systems	intrinsic	generalism	Number of organ systems the parasite is known to infect	#	Primary literature	79.6%	94.7%
7	Host Evolutionary Distinctiveness	host-related	generalism	Fair proportion measure of host evolutionary distinctiveness	million years (MY)	Dallas et al. 2018	92.5%	100.0%
8	Parasite Betweenness Centrality	bipartite network	generalism	A value describing the centrality of a species in the network by its position between other nodes (i.e., the number of shortest paths between all species that pass through the focal species)	NA	GMPD main	100.0%	100.0%

9	Host Betweenness Centrality	bipartite network	generalism	A value describing the centrality of a species in the network by its position between other nodes (i.e., the number of shortest paths between all species that pass through the focal species)	NA	GMPD main	100.0%	100.0%
10	Number of Ecoregions	environment	generalism	Number of ecoregions the parasite has been found (according to GMPD coordinates)	#	WFF TEOW	99.1%	100.0%
11	Proportion of Community Zoonotic	community	community	Proportion of species in the parasite's community that are zoonotic	%	GMPD main	100.0%	100.0%
12	Proportion of Hosts Zoonotic Carriers	community	community	Proportion of the parasite's hosts that are associated with a zoonotic parasite	%	GMPD main	100.0%	100.0%
13	Parasite Closeness Centrality	bipartite network	community	A value describing the centrality of a species in the network by its path lengths to other nodes (i.e., the relative distance from the focal species to all other species)	NA	GMPD main	100.0%	100.0%

14	Has Domestic Host	intrinsic	community	binary variable indicating if parasite species has a domestic host or not	binary	Primary literature	47.3%	36.8%
15	Host Proportional Generality	bipartite network	community	The number of associated parasites relative to the number of possible parasite associations.	NA	GMPD main	100.0%	100.0%
16	Class	intrinsic	intrinsic	Taxonomic class	factor	GMPD taxonomy	100.0%	100.0%
17	Family	intrinsic	intrinsic	Taxonomic family	factor	GMPD taxonomy	100.0%	100.0%
18	Intracellular	intrinsic	intrinsic	binary variable indicating if parasite species is intracellular or not	binary	Primary literature	79.2%	94.7%
19	Close Contact Transmission	intrinsic	intrinsic	Transmissible via close non-sexual contact such as grooming, biting, scratching, aerosols.	binary	GMPD trait	77.0%	94.7%
20	Vector Transmission	intrinsic	intrinsic	Transmissible by biting arthropod vectors.	binary	GMPD trait	77.0%	94.7%
21	Host Interbirth Interval	host-related	hosts	The length of time between successive births of the same female(s)	days	PanTHERIA	96.9%	100.0%
22	Host Trophic Level	host-related	hosts	Three factors corresponding to	factor	PanTHERIA	96.9%	94.7%

				trophic levels for each host species: herbivore; omnivore; carnivore				
23	Host Population Density	host-related	hosts	Number of individuals per square kilometer	#/km^2	PanTHERIA	93.8%	100.0%
24	Host Diet Invertebrate	host-related	hosts	% of host diet comprised by invertebrates	%	PHYLACINE	98.7%	100.0%
25	Primary Biogeographic Region	environment	geographic	The continental-scale biogeographic regions defined by differences in geologic and climatic history that contain distinct assemblages of plants and animals	factor	WFF TEOW	83.2%	94.7%
26	Mean Human Population Density	host-related	geographic	Mean human population density (persons per km2)	#/km^2	PanTHERIA	94.7%	100.0%
27	Mean Temperature	host-related	geographic	Mean monthly temperature (0.1°C)	degrees C	PanTHERIA	94.7%	100.0%
28	Mean Precipitation	host-related	geographic	Mean monthly precipitation (mm)	mm	PanTHERIA	94.7%	100.0%
29	Number of Web of Science Citations	study effort	sampling	Number of Web of Science hits for each protozoa species as of February 6, 2017	#	GMPD main	99.1%	94.7%

Table A2. BRT model parameter values

Parameter	Definition	Value
eta (η)	Step size shrinkage used to shrink the variable weights after each boosting step. Makes the boosting process more conservative to prevent overfitting. Default: 0.3	0.02
max.depth	Maximum depth of a tree. [more detail] Default: 6	3
alpha (α)	L1 regularization term on weights (analogous to Lasso regression). Increasing this value will make model more conservative. Default: 0	0.35
gamma (γ)	Minimum loss reduction required to make a further partition on a leaf node of the tree (a node is split only when the resulting split gives a positive reduction in the loss function). The larger gamma is, the more conservative the algorithm will be.	0.30
nrounds	Maximum number of boosting iterations.	64

APPENDIX B: RESULTS

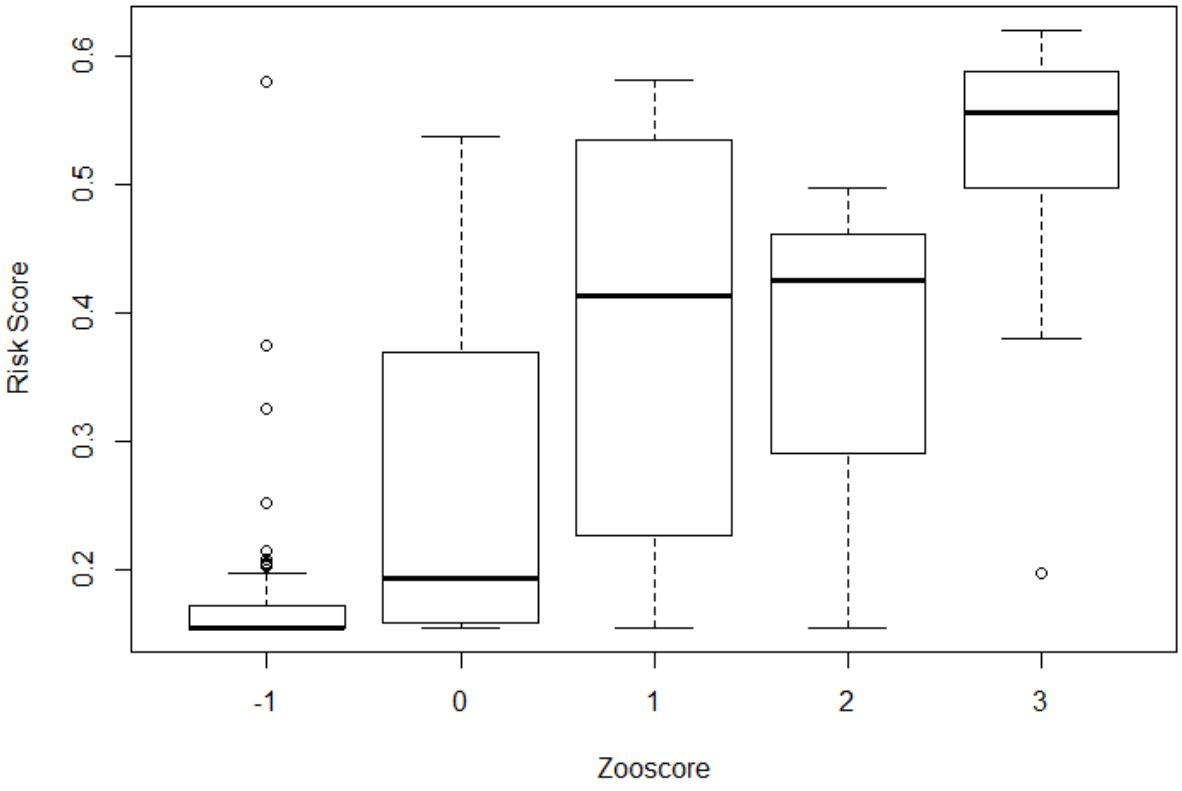


Figure A1. Box-and-whisker plots of risk scores for parasites of each zooscore. Each boxplot displays the median and the interquartile range (IQR). The extreme of the lower whisker is the minimum (Q1 - 1.5 * IQR), the extreme of the upper whisker is the maximum (Q3 + 1.5 * IQR), and the circles represent outliers.

Table A3. Confusion matrix for training data predictions

		Actual class: Non-zoonotic (0)	Actual class: Zoonotic (1)
Predicted class	Non-zoonotic (0)	124	0
	Zoonotic (1)	11	13

Table A4. Confusion matrix for testing data predictions

		Actual class: Non-zoonotic (0)	Actual class: Zoonotic (1)
Predicted class	Non-zoonotic (0)	62	2
	Zoonotic (1)	10	4

Table A5. Predicted probabilities for each protozoa species

Rank	Parasite Binomial Name	Zoonotic Status	Zooscore	Risk Score
1	*Entamoeba coli*	1	3	0.6196
2	*Trypanosoma brucei*	1	3	0.6034
3	*Toxoplasma gondii*	1	3	0.5880
4	*Entamoeba polecki*	1	1	0.5813
5	*Cryptosporidium parvum*	1	3	0.5805
6	*Neospora caninum*	0	-1	0.5805
7	*Leishmania infantum*	1	3	0.5565
8	*Entamoeba hartmanni*	1	1	0.5559
9	*Leishmania chagasi*	1	3	0.5461
10	*Entamoeba histolytica*	0	0	0.5370
11	*Entamoeba dispar*	0	0	0.5320
12	*Entamoeba chattoni*	1	1	0.5129
13	*Plasmodium brasilianum*	1	3	0.4968
14	*Plasmodium cynomolgi*	1	2	0.4968
15	*Leishmania donovani*	0	0	0.4962
16	*Leishmania braziliensis*	0	0	0.4688
17	*Balantidium coli*	1	1	0.4476
18	*Plasmodium inui*	1	2	0.4253
19	*Giardia intestinalis*	0	0	0.3929
20	*Leishmania shawi*	1	1	0.3805
21	*Trypanosoma cruzi*	1	3	0.3804
22	*Trypanosoma minasense*	0	-1	0.3745
23	*Plasmodium falciparum*	0	0	0.3691
24	*Plasmodium vivax*	0	0	0.3691
25	*Chilomastix mesnili*	0	0	0.3321
26	*Iodamoeba butschlii*	0	0	0.3321
27	*Trypanosoma conorhini*	0	-1	0.3254
28	*Besnoitia tarandi*	0	-1	0.2525
29	*Trypanosoma rangeli*	1	1	0.2493
30	*Sarcocystis neurona*	0	-1	0.2154
31	*Hepatozoon canis*	0	-1	0.2083
32	*Hepatocystis semnopitheci*	0	-1	0.2060
33	*Plasmodium gonderi*	0	-1	0.2060
34	*Cytauxzoon felis*	0	-1	0.2053
35	*Hepatozoon americanum*	0	-1	0.2053
36	*Sarcocystis felis*	0	-1	0.2053
37	*Sarcocystis fusiformis*	0	-1	0.2053
38	*Sarcocystis hominis*	1	1	0.2053
39	*Endolimax nana*	0	0	0.2043
40	*Eimeria pachylepyron*	0	-1	0.2041

41	*Isospora endocallimici*	0	-1	0.2041
42	*Plasmodium hylobati*	0	-1	0.2041
43	*Plasmodium simium*	0	-1	0.2030
44	*Trypanosoma mycetae*	0	-1	0.2030
45	*Blastocystis hominis*	0	0	0.2012
46	*Plasmodium coatneyi*	0	-1	0.1976
47	*Plasmodium knowlesi*	1	3	0.1976
48	*Trypanosoma diasi*	0	-1	0.1976
49	*Trypanosoma sanmartini*	0	-1	0.1976
50	*Trypanosoma venezuelense*	0	-1	0.1976
51	*Dientamoeba fragilis*	0	0	0.1945
52	*Eimeria kamoshika*	0	-1	0.1935
53	*Hepatocystis taiwanensis*	0	-1	0.1935
54	*Isospora arctopitheci*	0	-1	0.1935
55	*Isospora cebi*	0	-1	0.1935
56	*Isospora saimirae*	0	-1	0.1935
57	*Plasmodium eylesi*	0	-1	0.1935
58	*Plasmodium fieldi*	0	-1	0.1935
59	*Plasmodium jefferyi*	0	-1	0.1935
60	*Plasmodium pitheci*	0	-1	0.1935
61	*Plasmodium youngi*	0	-1	0.1935
62	*Trypanosoma advieri*	0	-1	0.1935
63	*Trypanosoma cyclops*	0	-1	0.1935
64	*Trypanosoma devei*	0	-1	0.1935
65	*Trypanosoma hippicum*	0	-1	0.1935
66	*Trypanosoma lambrechti*	0	-1	0.1935
67	*Trypanosoma lesourdi*	0	-1	0.1935
68	*Cycloposthium scutigerum*	0	-1	0.1911
69	*Theileria cervi*	0	-1	0.1897
70	*Trypanosoma evansi*	0	0	0.1868
71	*Trypanosoma congolense*	0	0	0.1835
72	*Trypanosoma perodictici*	0	-1	0.1835
73	*Sarcocystis capracanis*	0	-1	0.1746
74	*Balantidium aragaoi*	0	-1	0.1741
75	*Cyclospora cercopitheci*	0	-1	0.1741
76	*Cyclospora colobi*	0	-1	0.1741
77	*Cyclospora papionis*	0	-1	0.1741
78	*Retortamonas intestinalis*	0	0	0.1741
79	*Trichomonas intestinalis*	0	0	0.1741
80	*Troglodytella abrassarti*	0	-1	0.1741
81	*Hepatocystis kochi*	0	-1	0.1723
82	*Sarcocystis axicuonis*	0	-1	0.1708
83	*Sarcocystis elegans*	0	-1	0.1708

84	*Hepatocystis cercopitheci*	0	-1	0.1664
85	*Babesia propitheci*	0	-1	0.1648
86	*Cytauxzoon manul*	0	-1	0.1648
87	*Sarcocystis danzani*	0	-1	0.1648
88	*Sarcocystis mongolica*	0	-1	0.1648
89	*Babesia odocoilei*	0	-1	0.1618
90	*Entodinium bicornutum*	0	-1	0.1616
91	*Entodinium dubardi*	0	-1	0.1616
92	*Entodinium furca*	0	-1	0.1616
93	*Entodinium minimum*	0	-1	0.1616
94	*Entodinium rectangulatum*	0	-1	0.1616
95	*Epidinium bulbiferum*	0	-1	0.1616
96	*Epidinium gigas*	0	-1	0.1616
97	*Eudiplodinium maggi*	0	-1	0.1616
98	*Metadinium medium*	0	-1	0.1616
99	*Cystoisospora felis*	0	-1	0.1614
100	*Hepatozoon procyonis*	0	-1	0.1614
101	*Babesia bigemina*	0	-1	0.1593
102	*Babesia bovis*	0	-1	0.1593
103	*Babesia divergens*	0	0	0.1593
104	*Babesia microti*	0	0	0.1593
105	*Babesia ovis*	0	-1	0.1593
106	*Babesia rossi*	0	-1	0.1593
107	*Cystoisospora belli*	0	0	0.1593
108	*Cystoisospora canis*	0	-1	0.1593
109	*Eimeria bovis*	0	-1	0.1593
110	*Eimeria zuernii*	0	-1	0.1593
111	*Entodinium caudatum*	0	-1	0.1593
112	*Epidinium caudatum*	0	-1	0.1593
113	*Pentatrichomonas hominis*	0	0	0.1593
114	*Plasmodium fragile*	0	-1	0.1593
115	*Plasmodium reichenowi*	0	-1	0.1593
116	*Polyplastron multivesiculatum*	0	-1	0.1593
117	*Spirodinium equi*	0	-1	0.1593
118	*Theileria annae*	0	-1	0.1593
119	*Theileria bicornis*	0	-1	0.1593
120	*Theileria equi*	0	-1	0.1593
121	*Theileria mutans*	0	-1	0.1593
122	*Triadinium galea*	0	-1	0.1593
123	*Trypanosoma cervi*	0	-1	0.1593
124	*Trypanosoma primatum*	0	-1	0.1593
125	*Babesia bicornis*	0	-1	0.1558
126	*Babesia cellii*	0	-1	0.1558

127	*Babesia cynicti*	0	-1	0.1558
128	*Babesia EU2*	1	2	0.1558
129	*Babesia lotori*	0	-1	0.1558
130	*Babesia missirolii*	0	-1	0.1558
131	*Babesia pitheci*	0	-1	0.1558
132	*Blepharocorys jubata*	0	-1	0.1558
133	*Charonina hippopotami*	0	-1	0.1558
134	*Cycloposthium bipalmatum*	0	-1	0.1558
135	*Cycloposthium corrugatum*	0	-1	0.1558
136	*Cycloposthium dentiferum*	0	-1	0.1558
137	*Cycloposthium edentatum*	0	-1	0.1558
138	*Cyclospora cayetanensis*	0	0	0.1558
139	*Cystoisospora ohioensis*	0	-1	0.1558
140	*Diplodinium anacanthum*	0	-1	0.1558
141	*Diplodinium dentatum*	0	-1	0.1558
142	*Diploplastron affine*	0	-1	0.1558
143	*Eimeria auburnensis*	0	-1	0.1558
144	*Eimeria crispus*	0	-1	0.1558
145	*Eimeria felina*	0	-1	0.1558
146	*Eimeria furonis*	0	-1	0.1558
147	*Eimeria hreindyria*	0	-1	0.1558
148	*Eimeria ictidea*	0	-1	0.1558
149	*Eimeria macusaniensis*	0	-1	0.1558
150	*Eimeria madisonensis*	0	-1	0.1558
151	*Eimeria mayeri*	0	-1	0.1558
152	*Eimeria mccordocki*	0	-1	0.1558
153	*Eimeria melis*	0	-1	0.1558
154	*Eimeria mephitidis*	0	-1	0.1558
155	*Eimeria nuttalli*	0	-1	0.1558
156	*Eimeria odocoilei*	0	-1	0.1558
157	*Eimeria rangiferis*	0	-1	0.1558
158	*Eimeria serowi*	0	-1	0.1558
159	*Eimeria vulpis*	0	-1	0.1558
160	*Elyplastron bubali*	0	-1	0.1558
161	*Entodinium bursa*	0	-1	0.1558
162	*Entopolypoides macaci*	0	-1	0.1558
163	*Hepatocystis foleyi*	0	-1	0.1558
164	*Hepatocystis simiae*	0	-1	0.1558
165	*Isospora lutrae*	0	-1	0.1558
166	*Isospora melis*	0	-1	0.1558
167	*Isospora papionis*	0	-1	0.1558
168	*Isospora sengeri*	0	-1	0.1558
169	*Isospora spilogales*	0	-1	0.1558

170	*Isospora vulpis*	0	-1	0.1558
171	*Ostracodinium gracile*	0	-1	0.1558
172	*Ostracodinium mammosum*	0	-1	0.1558
173	*Ostracodinium trivesiculatum*	0	-1	0.1558
174	*Paraisotricha minuta*	0	-1	0.1558
175	*Paraplagiopyla kiboko*	0	-1	0.1558
176	*Parentodinium africanum*	0	-1	0.1558
177	*Parentodinium ostrea*	0	-1	0.1558
178	*Plasmodium gaboni*	0	-1	0.1558
179	*Plasmodium georgesi*	0	-1	0.1558
180	*Plasmodium girardi*	0	-1	0.1558
181	*Plasmodium petersi*	0	-1	0.1558
182	*Plasmodium rodhaini*	0	0	0.1558
183	*Plasmodium schwetzi*	0	-1	0.1558
184	*Plasmodium shortii*	0	-1	0.1558
185	*Plasmodium simiovale*	0	0	0.1558
186	*Prototapirella gorillae*	0	-1	0.1558
187	*Sarcocystis alces*	0	-1	0.1558
188	*Sarcocystis alceslatrans*	0	-1	0.1558
189	*Sarcocystis americana*	0	-1	0.1558
190	*Sarcocystis arctosi*	0	-1	0.1558
191	*Sarcocystis arieticanis*	0	-1	0.1558
192	*Sarcocystis camelopardalis*	0	-1	0.1558
193	*Sarcocystis capreolicanis*	0	-1	0.1558
194	*Sarcocystis cornagliai*	0	-1	0.1558
195	*Sarcocystis dubeyella*	0	-1	0.1558
196	*Sarcocystis gazellae*	0	-1	0.1558
197	*Sarcocystis giraffae*	0	-1	0.1558
198	*Sarcocystis grueneri*	0	-1	0.1558
199	*Sarcocystis hardangeri*	0	-1	0.1558
200	*Sarcocystis hemioni*	0	-1	0.1558
201	*Sarcocystis hemionilatrantis*	0	-1	0.1558
202	*Sarcocystis kirkpatricki*	0	-1	0.1558
203	*Sarcocystis klaseriensis*	0	-1	0.1558
204	*Sarcocystis melampi*	0	-1	0.1558
205	*Sarcocystis mephitisi*	0	-1	0.1558
206	*Sarcocystis odocoileocanis*	0	-1	0.1558
207	*Sarcocystis ovalis*	0	-1	0.1558
208	*Sarcocystis phacochoeri*	0	-1	0.1558
209	*Sarcocystis rangi*	0	-1	0.1558
210	*Sarcocystis scandinavica*	0	-1	0.1558
211	*Sarcocystis sebeki*	0	-1	0.1558
212	*Sarcocystis silva*	0	-1	0.1558

213	*Sarcocystis tarandivulpes*	0	-1	0.1558
214	*Sarcocystis ursusi*	0	-1	0.1558
215	*Sarcocystis woodhousei*	0	-1	0.1558
216	*Sarcocystis youngi*	0	-1	0.1558
217	*Tetratoxum parvum*	0	-1	0.1558
218	*Tetratoxum unifasciculatum*	0	-1	0.1558
219	*Theileria buffeli*	0	-1	0.1558
220	*Theileria parva*	0	-1	0.1558
221	*Troglocorys cava*	0	-1	0.1558
222	*Trypanosoma irangiense*	0	-1	0.1558
223	*Trypanosoma pestanai*	0	-1	0.1558
224	*Trypanosoma simiae*	0	-1	0.1558
225	*Trypanosoma theileri*	0	-1	0.1558
226	*Trypanosoma vivax*	0	0	0.1558