Characterization of eukaryotic endosymbiont communities: uncovering the "parasitome"

By

Leah A. Owens

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
(Comparative Biomedical Sciences)

at the
UNIVERSITY OF WISCONSIN-MADISON
2023

Date of final oral examination: 08/26/2022

The dissertation is approved by the following members of the Final Oral Committee:
Tony L. Goldberg, Professor, Pathobiological Sciences
Laura J. Knoll, Professor, Medical Microbiology and Immunology
Lyric C. Bartholomay, Professor, Pathobiological Sciences
Mostafa Zamanian, Assistant Professor, Pathobiological Sciences
Johanna R. Elfenbein, Assistant Professor, Pathobiological Sciences

ABSTRACT

Endosymbionts (organisms residing inside a host) include bacteria, archaea, viruses, fungi, and non-fungal eukaryotes. Evidence is mounting that endosymbionts impact host health regardless of their place on the continuum from frank pathogens to beneficials. DNA sequencing based methods have enabled cultivation-free, high-throughput characterization of bacterial and fungal endosymbiont communities (e.g. the “microbiome”) leading to novel insights into microbial ecology and host health. The work presented in this dissertation details the development and application of new metabarcoding-based methods to characterize eukaryotic endosymbiont communities. Chapter one describes the use of DNA sequencing-based diagnostics for investigating an outbreak of lethal disease in chimpanzees in Sierra Leone, the discovery of a bacterial pathogen linked with mortalities, and the characterization of an unexpectedly diverse parasite community in these animals using previously published metabarcoding methods. Chapter two details the development of a system for parasite enrichment in clinical samples by removal of host material using CRISPR-Cas9, which increases the sensitivity of parasite detection by 75 %. Chapter three describes the development of a new, optimized “universal” metabarcoding-based parasite detection method. Vertebrate Eukaryotic endoSymbiont and Parasite Analysis, or VESPA, is used for identifying and enumerating all classes of non-fungal eukaryotic endosymbionts, including protozoa, helminths, and microsporidia associated with vertebrate hosts. When directly compared, VESPA identifies a greater prevalence and diversity of endosymbionts than microscopic evaluation. Overall, this work advances the study of molecular parasitology by doing for eukaryotic endosymbionts what has been done so impactfully for bacteria and fungi via microbiome analysis. I hope this work will lead to significant advances in our understanding of the ecology, evolution, and health impacts of eukaryotic endosymbiont communities.

DEDICATION

To Zach,

for doing a lot.

ACKNOWLEDGEMENTS

I gratefully acknowledge and thank:

Tony Goldberg

Laura Knoll

Lyric Bartholomay

Mostafa Zamanian

Johanna Elfenbein

Sagan Friant

Bruno Martorelli Di Genova

Barbara Colitti

Christopher Dunn

Tricia Fry

The Goldberg Lab at University of Wisconsin-Madison

The Parfrey Lab at University of British Columbia

Tacugama Chimpanzee Sanctuary in Sierra Leone staff and animal care team

Jason Brenner

Dale Bjorling

Jenny Dahlberg

Amy Krohn

Maggie Dorsman

DeAnna Theiss

Harold Owens, Patricia Owens, Iris Neuman, April Owens, Heather Owens, Sarah Owens, Carol Renthe, Dan Smith, Adam Smith

UW Madison Parasitology and Vector Biology Training Grant

UW Madison Global Health Institute Graduate Student Fellowship

TABLE OF CONTENTS

CHAPTER 1: Introduction

Symbiont communities

Our collective understanding of health and disease has shifted from a model of microorganisms as solely pathogenic agents to a more nuanced "holobiont" framework in which hosts and the organisms living in, on, and around them form discrete, functional ecosystems [1]. Symbionts (species residing in or on a host) colonize exterior surfaces [2], the entire lining of the gastrointestinal (GI) tract from oral cavity to anus [3], and are found in other tissues such as the liver [4], respiratory tract components [5], and urogenital organs [6]. Animals (including humans) are hosts to a wide range of symbiont biodiversity that spans the taxonomic tree of life, including viruses, bacteria, archaea, algae, fungi, microscopic protozoa, and macroscopic eukaryotes (reviewed in [7]).

Symbiont roles with regard to host function are equally varied. At one end of the spectrum are frank pathogens such as the ebolaviruses (*Filoviridae*: *Ebolavirus*), which can cause lethal disease with case fatality rates exceeding 80 % depending on viral species/strain, host factors, and reporting parameters [8]. At the other end of the spectrum are beneficial organisms such as rumen microbes with which the animal host has co-evolved and is reliant upon (directly as a protein source and indirectly for breakdown of feed into volatile fatty acids, amongst other functions) [9]. In many cases the roles of symbionts are dynamic, alternating between pathogenic and benign depending on context. For example, the bacterium responsible for human cholera disease, *Vibrio cholera,* can be a normal part of GI flora, but if critical density is reached, becomes pathogenic via expression of toxin genes [10]. Host immune factors also modulate the effects of symbionts. For instance, major histocompatibility complex (MHC) molecules mediate adaptive immune responses in vertebrates. Variation in MHCs at the gene level influence the composition of gut microbial communities in individuals and, in turn, susceptibility to certain enteric infectious diseases [11].

In addition to interacting with members of the same species and immune system elements, symbionts also form mixed interspecies assemblages with important consequences. Host-associated organisms within communities interact: they communicate, cooperate, compete, and have emergent properties which directly impact host function in myriad ways (reviewed in [12]). For example, in animals with all types of digestive systems (ruminant, pseudo ruminant, monogastric, avian, etc.) gut microbial communities perform critical biochemical functions and impact everything from food animal production [13], to the endurance of racehorses [14], to drug metabolism [15], and behavior and welfare [16]. Knowledge of the mechanisms of such interactions is relatively new, but the empirical understanding of some of these relationships have had impacts on clinical medicine for centuries. For example, the content of a cow's rumen is so important to healthy GI function, that rumen fluid transplantation (known as transfaunation, referring to the microorganisms responsible for the therapeutic effect) has been used as a treatment for various clinical issues, including indigestion, in cattle since the 1700s [17].

Parasites as endosymbionts

One major category of symbionts important to human and animal medicine is parasites. Although diseases attributed to parasites kill millions of people every year [18], cause substantial morbidity [19], and disproportionately affect disadvantaged and high-risk populations [20], there is growing evidence that they also exist in an ecological balance with the host [21]. "Parasitic" organisms may also play an important, positive role in host health by aiding in digestion [22, 23], modulating protective immunity [24], favorably altering the bacterial microbiome [25] and, paradoxically, lowering risk of pathogenic parasite infection and disease [26]. For example, enteric helminths reduce clinical disease (*e.g.,* gastric atrophy) caused by *Helicobacter pylori* infection by modulating the gastric immune system [27] and chronic infection with the parasite *Toxoplasma gondii* prevents experimental cerebral malaria by upregulating systemic immune factors, including interleukins [28].

Herein I focus on and refer to eukaryotic endosymbionts, to include parasites and commensals, while differentiating ectoparasites (fleas, ticks, mites). Furthermore, I exclude fungal organisms from my focus based on their fundamentally transient association with hosts [29], but make an exception for microsporidia whose life cycles are considered closer to protozoa than other fungi [30].

Eukaryotic endosymbiont communities

There is mounting evidence that eukaryotic endosymbionts, including disease-associated parasites, form community assemblages important to “holobiont” function, in much the same way as bacteria and fungi (reviewed in [31]). The implications of such evidence in humans and animals are far-reaching. For example, it may be possible to modulate the eukaryotic symbiont community to improve health or disease outcomes, much as has become common practice for bacterial communities [32]. Similarly, broad-spectrum anti-parasitic drugs could inadvertently lead to dysbiosis and increased disease risk if their effects on symbiont communities are not understood [33].

Our current knowledge of eukaryotic symbionts is highly skewed towards the true parasites that cause disease; we still know very little about the structure and function of communities of protozoans and helminths that also colonize hosts. It is therefore essential that we study eukaryotic symbiont assemblages as we have bacterial assemblages, to understand the basic function of these communities, how they interact with other components of the microbiome, and how they mediate parasite infection, disease, and overall health. Recent studies have addressed aspects of eukaryotic endosymbiont community, but with limited success, due in large part to methodological limitations.

Methodological approaches for studying endosymbiont communities

Within a clinical framework, identification of host-associated organisms is focused on diagnosis and treatment. Clinical signs, history, and physical examination findings are used to create a differential diagnosis list which guides targeted testing for specific, suspected pathogenic organisms, ideally leading to a definitive diagnosis, treatment, and clinical resolution. In contrast, the study of assemblages of organisms and their collective properties requires alternative approaches.

In the case of parasites, although current standard methods are of great utility, they are largely designed to detect specific pathogens and do not translate well to the study of symbiont assemblages [34]. Identification by microscopy has been a gold-standard method since the 17th century [35] and remains an important diagnostic tool [36], but it is time-consuming, can be difficult to accurately compare across studies (due to variation in enrichment methods, subjectivity in assessment, etc.) [37], and cannot resolve certain organism complexes which are indistinguishable based on morphology alone [38]. More recent applications of antigen- and DNA-based diagnostics to parasites, such as rapid tests for *Dirofilaria immitis* antigen in blood [39] and Polymerase Chain Reaction (PCR) [40] and quantitative PCR (qPCR) tests for GI protozoa in feces [41], have led to greater sensitivity and resolution, but still require use within a limited framework since each assay is designed to detect one or a few known pathogens (reviewed in [42]).

In the study of symbiont assemblages, the goal is to identify the entire community of organisms, which requires approaches often borrowed from the field of ecology [43]. The first studies of mixed communities relied on growing representative organisms *in vitro*, which has limitations as many organisms are fastidious to culture [44]. The field has since largely shifted to culture-independent techniques based on massively parallel DNA sequencing which are sensitive (*i.e.* detect low abundance organisms), untargeted (*i.e.* identify all organisms of a particular type, regardless of the suspected composition of the sample), and high-throughput (*i.e.* assess many samples simultaneously) [45].

For viruses, and increasingly bacteria and archaea, the metagenome, the collective genetic material of all organisms present [46], can be fragmented and sequenced in its entirety using shotgun metagenomics. Due to the complexity and size of resulting data sets, metagenomic analysis, particularly of eukaryotic assemblages, is currently intractable in most settings [47]. Sequencing of amplified marker genes, or metagenomic barcoding, is a more widely used technique. Application of this approach to the bacterial 16S [48] and fungal ITS [49] genes have already revolutionized our understanding of complex communities and standardized protocols to study these organisms are available from sample preparation through analysis.

Metagenomic barcoding methods overview

Metagenomic barcoding (metabarcoding), also known as targeted amplicon sequencing, is a PCR-based method in which a specific area of the genome is amplified that serves as a marker or "barcode". Instead of sequencing a full genome from each organism, only the barcode undergoes sequencing and serves as a proxy which is used to identify the source organism (reviewed in [50]). Barcode genes must be carefully selected such that the external regions are well-conserved for placement of broad primers, but internal regions contain sufficient nucleotide sequence diversity to distinguish a range of organisms from each other [51]. Ribosomal RNA (rRNA) genes are particularly well-suited to this purpose. rRNA subunits consist of islands of conserved sequence corresponding to external surface features of the 3D molecule, interspersed with internal regions under less selective pressure, which are more divergent [52]. Primers must also be designed carefully to ensure sequence complementarity to all targeted groups without allowing for interfering off-target amplification [53].

As shown schematically in **Figure 1**, metabarcoding protocols begin with DNA extraction from samples followed by PCR which amplifies the selected barcode (reviewed in [54]). Resulting amplicons must be further processed to enable DNA sequencing by addition of unique indexes (required to later identify which data came from which samples *i.e.,* demultiplexing) and adapter sequences (required for physical binding to the surface of the sequencing machine), after which they are referred to as “sequencing libraries”. There are several methods of library preparation [55], but a technique commonly used in bacterial metabarcoding and featured in the work presented here utilizes a two-step PCR approach. The initial PCR uses compound primers consisting of a locus-specific sequence which binds the DNA region of interest, linked to an adapter sequence (also known as an “overhang”) which does not bind the template, but serves as the attachment site for primers used subsequent PCR (**Figure 2**). Amplicons from the initial PCR are “cleaned up” to remove excess reagents and then used as template for a second PCR reaction. The primers in the second reaction bind to the overhangs and themselves contain the unique indexes and adapters described above [56]. The products are purified once again to remove excess reagents and then subjected to DNA sequencing.

Eukaryotic endosymbiont metabarcoding

Culture-independent metabarcoding methods have been established and standardized in the contexts of bacterial and fungal microbiome research [57, 58], environmental biodiversity surveys [55], and animal dietary studies [59], but no universally-accepted method exists for eukaryotic endosymbionts. Several such studies have been published in recent years with promising results, but long-standing unresolved technical issues remain. Primer complementarity is the most important factor for metabarcoding quality [60, 61], determines the quantitative capacity of the assay [62], and is one of the most challenging parts of the assay to design. Thus far, published studies have used primers borrowed from environmental research [63, 64], based on limited sequence data [65], or designed to target only protozoans [66], metazoans [67], or helminths [68-70]. Complex genomes, like those of eukaryotes, have additional hurdles to metabarcoding assay design. For example, whereas bacterial rRNA genes generally exist in low copy number and roughly equal lengths, eukaryotes display a huge range of rRNA copy number variations and length polymorphisms [71].

The challenges of metabarcoding eukaryotic endosymbionts are clear when one considers all of the organisms that fall under the umbrella of “parasites”. A metabarcoding method must detect everything from the tiniest *Blastocystis,* to worms that are meters long, and whose rRNA genes are as diverse as their morphology [72]. Because “universal” metabarcoding primers target a large swath of eukaryotic organisms, they often recognize some, if not all, vertebrate host species as well. In some clinically-important sample types, animal and human DNA is highly abundant compared to endosymbiont DNA so host signal can bias results or mask target organisms completely, necessitating target enrichment. Some studies have used restriction enzyme digests of host sequences [73, 74] or host amplification blocking mechanisms [75, 76], but no one method has been shown to be consistently effective. Other studies have used higher sequencing coverage in lieu of a blocking mechanism [77], which is problematic because rare sequences cannot be reliably detected by increasing sequencing depth alone [55].

The methodological gap in endosymbiont barcoding truly needs to be filled in order to enable large-scale research into the basic biology and ecology of host-associated symbiont assemblages for the first time. This knowledge will impact future health management strategies and parasitic disease treatments and will contribute to progress in clinical parasitology and public health by informing development of unbiased tools for identifying parasite infections and co-infections across host populations and sample types.

Thesis synopsis

In the dissertation research presented herein, I examine host-associated organism assemblages in various contexts, identify issues with current methods of studying eukaryotic endosymbiont communities, and build upon existing techniques through the development of new tools to more accurately characterize such communities.

Chapter 2 describes the investigation of a lethal disease outbreak of unknown etiology in sanctuary chimpanzees in Sierra Leone. I use DNA sequencing-based techniques to identify case associated organisms of all classes, identify a bacterium which is strongly linked to the disease (termed epizootic neurologic and gastrointestinal syndrome, ENGS), and further characterize the bacterium as a new clostridial species, *Sarcina troglodytae*. As part of the investigation, I use a parasite metabarcoding method which identifies a diversity of endosymbionts in chimpanzee tissue and fecal samples. Although no parasites are case-associated and thus not pursued further as potential disease-causing agents, the experiments bring to light the drawbacks of published methods for endosymbiont metabarcoding. There is a high abundance of off target signal of both prokaryotic and host origin, demonstrating the need for enrichment, along with a dearth of reads from certain groups of parasites, demonstrating the need to re-visit primer coverage.

In **Chapter 3,** I investigate the issue of host signal in eukaryotic endosymbiont metabarcoding protocols. I demonstrate the need for enrichment of target sequences in host-dense sample types, the lack of efficacy of some published host signal reduction methods and develop a new approach using CRISPR-Cas9 cleavage directed by host-specific guideRNAs. The addition of CRISPR-Cas9 enrichment increases diagnostic sensitivity by nearly 76 % in experimentally infected animals. Furthermore, this enrichment method allows for metabarcoding detection of natural infections in hosts that are otherwise undetectable.

The most fundamental challenges to endosymbiont metabarcoding (*i.e.,* primer design and taxonomic coverage breadth) are the focus of **Chapter 4** in which I test commonly published metabarcoding protocols and find widespread issues with primer coverage. After a thorough literature review, I undertake the design of a new metabarcoding method, including the design of pan-endosymbiont primers, *in silico* comparisons to published primers, testing with a parasite community standard, and finalization of a new metabarcoding pipeline protocol, which I call VESPA (Vertebrate Eukaryotic endoSymbiont and Parasite Analysis). Finally, I apply VESPA to clinical samples and directly compare findings made by microscopic examination to metabarcoding results. Overall metabarcoding successfully recapitulates microscopy findings and identifies a greater diversity of organisms in a greater number of individuals.

In **Chapter 5,** I conclude with a brief summary of the results of this research, identify some of the major impacts of my findings, and identify directions for future work. Specifically, I address the need for further study of the *Sarcina* bacterium associated with the outbreak of ENGS in Sierra Leone and the many exciting applications of VESPA and CRISPR-Cas9 host signal reduction.

References

1. Bordenstein SR, Theis KR. Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. PLoS Biol. 2015;13(8):e1002226. Epub 2015/08/19. doi: 10.1371/journal.pbio.1002226. PubMed PMID: 26284777; PubMed Central PMCID: PMCPMC4540581.

2. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, et al. Topographical and temporal diversity of the human skin microbiome. Science. 2009;324(5931):1190-2. Epub 2009/05/30. doi: 10.1126/science.1171700. PubMed PMID: 19478181; PubMed Central PMCID: PMCPMC2805064.

3. Yasuda K, Oh K, Ren B, Tickle TL, Franzosa EA, Wachtman LM, et al. Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. Cell Host Microbe. 2015;17(3):385-91. Epub 2015/03/04. doi: 10.1016/j.chom.2015.01.015. PubMed PMID: 25732063; PubMed Central PMCID: PMCPMC4369771.

4. Balmer ML, Slack E, de Gottardi A, Lawson MA, Hapfelmeier S, Miele L, et al. The liver may act as a firewall mediating mutualism between the host and its gut commensal microbiota. Sci Transl Med. 2014;6(237):237ra66. Epub 2014/05/23. doi: 10.1126/scitranslmed.3008618. PubMed PMID: 24848256.

5. Bassis CM, Erb-Downward JR, Dickson RP, Freeman CM, Schmidt TM, Young VB, et al. Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. mBio. 2015;6(2):e00037. Epub 2015/03/05. doi: 10.1128/mBio.00037-15. PubMed PMID: 25736890; PubMed Central PMCID: PMCPMC4358017.

6. Thomas-White K, Forster SC, Kumar N, Van Kuiken M, Putonti C, Stares MD, et al. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. Nat Commun. 2018;9(1):1557. Epub 2018/04/21. doi: 10.1038/s41467-018-03968-5. PubMed PMID: 29674608; PubMed Central PMCID: PMCPMC5908796.

7. Douglas AE. Housing microbial symbionts: evolutionary origins and diversification of symbiotic organs in animals. Philos Trans R Soc Lond B Biol Sci. 2020;375(1808):20190603. Epub 2020/08/11. doi: 10.1098/rstb.2019.0603. PubMed PMID: 32772661; PubMed Central PMCID: PMCPMC7435165.

8. Garske T, Cori A, Ariyarajah A, Blake IM, Dorigatti I, Eckmanns T, et al. Heterogeneities in the case fatality ratio in the West African Ebola outbreak 2013-2016. Philos Trans R Soc Lond B Biol Sci. 2017;372(1721). Epub 2017/04/12. doi: 10.1098/rstb.2016.0308. PubMed PMID: 28396479; PubMed Central PMCID: PMCPMC5394646.

9. WILLIAMS A. GS COLEMAN. The Rumen Microbial Ecosystem. 1997:73.

10. Holmgren J. Actions of cholera toxin and the prevention and treatment of cholera. Nature. 1981;292(5822):413-17. Epub 1981/07/30. doi: 10.1038/292413a0. PubMed PMID: 7019725.

11. Kubinak JL, Stephens WZ, Soto R, Petersen C, Chiaro T, Gogokhia L, et al. MHC variation sculpts individualized microbial communities that control susceptibility to enteric infection. Nat Commun. 2015;6:8642. Epub 2015/10/27. doi: 10.1038/ncomms9642. PubMed PMID: 26494419; PubMed Central PMCID: PMCPMC4621775.

12. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24(4):392-400. Epub 2018/04/11. doi: 10.1038/nm.4517. PubMed PMID: 29634682; PubMed Central PMCID: PMCPMC7043356.

13. Kogut MH, Arsenault RJ. Editorial: Gut Health: The New Paradigm in Food Animal Production. Front Vet Sci. 2016;3:71. Epub 2016/09/16. doi: 10.3389/fvets.2016.00071. PubMed PMID: 27630994; PubMed Central PMCID: PMCPMC5005397.

14. Plancade S, Clark A, Philippe C, Helbling JC, Moisan MP, Esquerre D, et al. Unraveling the effects of the gut microbiota composition and function on horse endurance physiology. Sci Rep. 2019;9(1):9620. Epub 2019/07/05. doi: 10.1038/s41598-019-46118-7. PubMed PMID: 31270376; PubMed Central PMCID: PMCPMC6610142.

15. Javdan B, Lopez JG, Chankhamjon P, Lee YJ, Hull R, Wu Q, et al. Personalized Mapping of Drug Metabolism by the Human Gut Microbiome. Cell. 2020;181(7):1661-79 e22. Epub 2020/06/12. doi: 10.1016/j.cell.2020.05.001. PubMed PMID: 32526207; PubMed Central PMCID: PMCPMC8591631.

16. Kraimi N, Dawkins M, Gebhardt-Henrich SG, Velge P, Rychlik I, Volf J, et al. Influence of the microbiota-gut-brain axis on behavior and welfare in farm animals: A review. Physiol Behav. 2019;210:112658. Epub 2019/08/21. doi: 10.1016/j.physbeh.2019.112658. PubMed PMID: 31430443.

17. Hungate RE. The rumen and its microbes: Elsevier; 2013.

18. Collaborators GBDCoD. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736-88. Epub 2018/11/30. doi: 10.1016/S0140-6736(18)32203-7. PubMed PMID: 30496103; PubMed Central PMCID: PMCPMC6227606.

19. DALYs GBD, Collaborators H. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1859-922. Epub 2018/11/13. doi:

10.1016/S0140-6736(18)32335-3. PubMed PMID: 30415748; PubMed Central PMCID: PMCPMC6252083.

20. Hotez PJ, Alvarado M, Basanez MG, Bolliger I, Bourne R, Boussinesq M, et al. The global burden of disease study 2010: interpretation and implications for the neglected tropical diseases. PLoS Negl Trop Dis. 2014;8(7):e2865. Epub 2014/07/25. doi: 10.1371/journal.pntd.0002865. PubMed PMID: 25058013; PubMed Central PMCID: PMCPMC4109880.

21. Lukes J, Stensvold CR, Jirku-Pomajbikova K, Parfrey LW. Are Human Intestinal Eukaryotes Beneficial or Commensals? PLoS Path. 2015;11(8). doi: 10.1371/journal.ppat.1005039. PubMed PMID: 26270819. PMCID: PMC4536199.

22. Profousova I, Mihalikova K, Laho T, Varadyova Z, Petrzelkova KJ, Modry D, et al. The ciliate, *Troglodytella abrassarti*, contributes to polysaccharide hydrolytic activities in the chimpanzee colon. Folia Microbiol. 2011;56(4):339-43. doi: 10.1007/s12223-011-0053-x. PubMed PMID: 21818613.

23. Takenaka A, Tajima K, Mitsumori M, Kajikawa H. Fiber digestion by rumen ciliate protozoa. Microbes Environ. 2004;19(3):203-10. PubMed PMID: Zoorec:Zoor14102014360.

24. Chudnovskiy A, Mortha A, Kana V, Kennard A, Ramirez JD, Rahman A, et al. Host-Protozoan Interactions Protect from Mucosal Infections through Activation of the Inflammasome. Cell. 2016;167(2):444-56 e14. Epub 2016/10/08. doi: 10.1016/j.cell.2016.08.076. PubMed PMID: 27716507; PubMed Central PMCID: PMCPMC5129837.

25. Audebert C, Even G, Cian A, Blastocystis Investigation G, Loywick A, Merlin S, et al. Colonization with the enteric protozoa *Blastocystis* is associated with increased diversity of human gut bacterial microbiota. Sci Rep. 2016;6:25255. Epub 2016/05/06. doi: 10.1038/srep25255. PubMed PMID: 27147260; PubMed Central PMCID: PMCPMC4857090.

26. Ashby B, King KC. Friendly foes: The evolution of host protection by a parasite. Evol Lett. 2017;1(4):211-21. Epub 2017/08/31. doi: 10.1002/evl3.19. PubMed PMID: 30283650; PubMed Central PMCID: PMCPMC6121858.

27. Fox JG, Beck P, Dangler CA, Whary MT, Wang TC, Shi HN, et al. Concurrent enteric helminth infection modulates inflammation and gastric immune responses and reduces helicobacter-induced gastric atrophy. Nat Med. 2000;6(5):536-42. Epub 2000/05/10. doi: 10.1038/75015. PubMed PMID: 10802709.

28. Settles EW, Moser LA, Harris TH, Knoll LJ. *Toxoplasma gondii* upregulates interleukin-12 to prevent Plasmodium berghei-induced experimental cerebral malaria. Infect Immun.

2014;82(3):1343-53. Epub 2014/01/08. doi: 10.1128/IAI.01259-13. PubMed PMID: 24396042; PubMed Central PMCID: PMCPMC3957979.

29. Kohler JR, Hube B, Puccia R, Casadevall A, Perfect JR. Fungi that Infect Humans. Microbiol Spectr. 2017;5(3). Epub 2017/06/10. doi: 10.1128/microbiolspec.FUNK-0014-2016. PubMed PMID: 28597822.

30. Vossbrinck CR, Debrunner-Vossbrinck BA. Molecular phylogeny of the Microsporidia: ecological, ultrastructural and taxonomic considerations. Folia Parasitol (Praha). 2005;52(1-2):131-42; discussion 0. Epub 2005/07/12. doi: 10.14411/fp.2005.017. PubMed PMID: 16004372.

31. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258-70.

32. Chehri M, Christensen AH, Halkjaer SI, Gunther S, Petersen AM, Helms M. Case series of successful treatment with fecal microbiota transplant (FMT) oral capsules mixed from multiple donors even in patients previously treated with FMT enemas for recurrent Clostridium difficile infection. Medicine. 2018;97(31). doi: 10.1097/MD.0000000000011706. PubMed PMCID: PMC6081131.

33. Leung JM, Graham AL, Knowles SCL. Parasite-Microbiota Interactions With the Vertebrate Gut: Synthesis Through an Ecological Lens. Front Microbiol. 2018;9:843. Epub 2018/06/06. doi: 10.3389/fmicb.2018.00843. PubMed PMID: 29867790; PubMed Central PMCID: PMCPMC5960673.

34. Laforest-Lapointe I, Arrieta MC. Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. mSystems. 2018;3(2). Epub 2018/03/21. doi: 10.1128/mSystems.00201-17. PubMed PMID: 29556538; PubMed Central PMCID: PMCPMC5850078.

35. Dobell C. The Discovery of the Intestinal Protozoa of Man. Proc R Soc Med. 1920;13(Sect Hist Med):1-15. Epub 1920/01/01. PubMed PMID: 19981292; PubMed Central PMCID: PMCPMC2151982.

36. Leal SM, Jr., Rodino KG, Fowler WC, Gilligan PH. Practical Guidance for Clinical Microbiology Laboratories: Diagnosis of Ocular Infections. Clin Microbiol Rev. 2021;34(3):e0007019. Epub 2021/06/03. doi: 10.1128/CMR.00070-19. PubMed PMID: 34076493; PubMed Central PMCID: PMCPMC8262805.

37. Libman MD, Gyorkos TW, Kokoskin E, Maclean JD. Detection of pathogenic protozoa in the diagnostic laboratory: result reproducibility, specimen pooling, and competency assessment. J Clin Microbiol. 2008;46(7):2200-5. Epub 2008/05/02. doi: 10.1128/JCM.01666-07. PubMed PMID: 18448690; PubMed Central PMCID: PMCPMC2446938.

38. Nadler SA, GP DEL. Integrating molecular and morphological approaches for characterizing parasite cryptic species: implications for parasitology. Parasitology. 2011;138(13):1688-709. Epub 2011/02/02. doi: 10.1017/S003118201000168X. PubMed PMID: 21281559.

39. Panarese R, Iatta R, Mendoza-Roldan JA, Szlosek D, Braff J, Liu J, et al. Comparison of Diagnostic Tools for the Detection of *Dirofilaria immitis* Infection in Dogs. Pathogens. 2020;9(6). Epub 2020/06/26. doi: 10.3390/pathogens9060499. PubMed PMID: 32580453; PubMed Central PMCID: PMCPMC7350293.

40. Autier B, Gangneux JP, Robert-Gangneux F. Evaluation of the Allplex(TM) Gastrointestinal Panel-Parasite Assay for Protozoa Detection in Stool Samples: A Retrospective and Prospective Study. Microorganisms. 2020;8(4). Epub 2020/04/25. doi: 10.3390/microorganisms8040569. PubMed PMID: 32326453; PubMed Central PMCID: PMCPMC7232139.

41. Menu E, Mary C, Toga I, Raoult D, Ranque S, Bittar F. A hospital qPCR-based survey of 10 gastrointestinal parasites in routine diagnostic screening, Marseille, France. Epidemiol Infect. 2019;147:e100. Epub 2019/03/15. doi: 10.1017/S0950268819000165. PubMed PMID: 30869032; PubMed Central PMCID: PMCPMC6518462.

42. Kumar S, Gupta S, Mohmad A, Fular A, Parthasarathi BC, Chaubey AK. Molecular tools-advances, opportunities and prospects for the control of parasites of veterinary importance. Int J Trop Insect Sci. 2021;41(1):33-42. Epub 2020/08/25. doi: 10.1007/s42690-020-00213-9. PubMed PMID: 32837530; PubMed Central PMCID: PMCPMC7387080.

43. Bass D, Stentiford GD, Littlewood DTJ, Hartikainen H. Diverse Applications of Environmental DNA Methods in Parasitology. Trends Parasitol. 2015;31(10):499-513. Epub 2015/10/05. doi: 10.1016/j.pt.2015.06.013. PubMed PMID: 26433253.

44. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems. 2018;3(5). Epub 2018/10/03. doi: 10.1128/mSystems.00055-18. PubMed PMID: 30273414; PubMed Central PMCID: PMCPMC6156271.

45. Moore RA, Warren RL, Freeman JD, Gustavsen JA, Chenard C, Friedman JM, et al. The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PLoS One. 2011;6(5):e19838. Epub 2011/05/24. doi: 10.1371/journal.pone.0019838. PubMed PMID: 21603639; PubMed Central PMCID: PMCPMC3094400.

46. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833-44. Epub 2017/09/13. doi: 10.1038/nbt.3935. PubMed PMID: 28898207.

47. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. J Infect Dis. 2020;221(Suppl 3):S292-S307. Epub 2019/10/16. doi: 10.1093/infdis/jiz286. PubMed PMID: 31612214.

48. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, et al. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics. 2016;17:55. Epub 2016/01/15. doi: 10.1186/s12864-015-2194-9. PubMed PMID: 26763898; PubMed Central PMCID: PMCPMC4712552.

49. Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L. Mycobiome diversity: high-throughput sequencing and identification of fungi. Nat Rev Microbiol. 2019;17(2):95-109. Epub 2018/11/18. doi: 10.1038/s41579-018-0116-y. PubMed PMID: 30442909.

50. Cristescu ME. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol Evol. 2014;29(10):566-71. Epub 2014/09/02. doi: 10.1016/j.tree.2014.08.001. PubMed PMID: 25175416.

51. van der Loos LM, Nijland R. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Mol Ecol. 2020. Epub 2020/08/12. doi: 10.1111/mec.15592. PubMed PMID: 32779312.

52. Bradley IM, Pinto AJ, Guest JS. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. Appl Environ Microbiol. 2016;82(19):5878-91. doi: 10.1128/Aem.01630-16. PubMed PMID: 27451454 PMCID: PMC5038042.

53. Bedarf JR, Beraza N, Khazneh H, Ozkurt E, Baker D, Borger V, et al. Much ado about nothing? Off-target amplification can lead to false-positive bacterial brain microbiome detection in healthy and Parkinson's disease individuals. Microbiome. 2021;9(1):75. Epub 2021/03/28. doi: 10.1186/s40168-021-01012-1. PubMed PMID: 33771222; PubMed Central PMCID: PMCPMC8004470.

54. Bohmann K, Elbrecht V, Caroe C, Bista I, Leese F, Bunce M, et al. Strategies for sample labelling and library preparation in DNA metabarcoding studies. Mol Ecol Resour. 2022;22(4):1231-46. Epub 2021/09/23. doi: 10.1111/1755-0998.13512. PubMed PMID: 34551203; PubMed Central PMCID: PMCPMC9293284.

55. Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol Evol. 2018;9(1):134-47.

56. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010(6):pdb prot5448. Epub 2010/06/03. doi: 10.1101/pdb.prot5448. PubMed PMID: 20516186.

57. Kameoka S, Motooka D, Watanabe S, Kubo R, Jung N, Midorikawa Y, et al. Benchmark of 16S rRNA gene amplicon sequencing using Japanese gut microbiome data from the V1-V2 and V3-V4 primer sets. BMC Genomics. 2021;22(1):527. Epub 2021/07/12. doi: 10.1186/s12864-021-07746-4. PubMed PMID: 34246242; PubMed Central PMCID: PMCPMC8272389.

58. Op De Beeck M, Lievens B, Busschaert P, Declerck S, Vangronsveld J, Colpaert JV. Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. PLoS One. 2014;9(6):e97629. Epub 2014/06/17. doi: 10.1371/journal.pone.0097629. PubMed PMID: 24933453; PubMed Central PMCID: PMCPMC4059633.

59. De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, et al. DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol Ecol Resour. 2014;14(2):306-23. Epub 2013/10/17. doi: 10.1111/1755-0998.12188. PubMed PMID: 24128180.

60. Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, et al. Minimizing polymerase biases in metabarcoding. Mol Ecol Resour. 2018. Epub 2018/05/26. doi: 10.1111/1755-0998.12895. PubMed PMID: 29797549.

61. Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biol Lett. 2014;10(9). Epub 2014/09/12. doi: 10.1098/rsbl.2014.0562. PubMed PMID: 25209199; PubMed Central PMCID: PMCPMC4190964.

62. Pinol J, Mir G, Gomez-Polo P, Agusti N. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Mol Ecol Resour. 2015;15(4):819-30. Epub 2014/12/03. doi: 10.1111/1755-0998.12355. PubMed PMID: 25454249.

63. Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, et al. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. Frontiers in Microbiology. 2014;5. doi: 10.3389/fmicb.2014.00298. PubMed PMID: 24995004 PMCID: PMC4063188.

64. Mann AE, Mazel F, Lemay MA, Morien E, Billy V, Kowalewski M, et al. Biodiversity of protists and nematodes in the wild nonhuman primate gut. Isme J. 2020;14(2):609-22. Epub 2019/11/14. doi: 10.1038/s41396-019-0551-4. PubMed PMID: 31719654; PubMed Central PMCID: PMCPMC6976604.

65. Hugerth LW, Muller EEL, Hu YOO, Lebrun LAM, Roume H, Lundin D, et al. Systematic Design of 18S rRNA Gene Primers for Determining Eukaryotic Diversity in Microbial Consortia. PLoS One. 2014;9(4). doi: 10.1371/journal.pone.0095567. PubMed PMID: 24755918 PMCID: PMC3995771.

66. Del Campo J, Pons MJ, Herranz M, Wakeman KC, Del Valle J, Vermeij MJA, et al. Validation of a universal set of primers to study animal-associated microeukaryotic communities. Environ Microbiol. 2019;21(10):3855-61. Epub 2019/07/07. doi: 10.1111/1462-2920.14733. PubMed PMID: 31278828.

67. Wood JR, Díaz FP, Latorre C, Wilmshurst JM, Burge OR, González F, et al. Ancient parasite DNA from late Quaternary Atacama Desert rodent middens. Quat Sci Rev. 2019;226:106031.

68. Aivelo T, Harris K, Cadle J, Wright P. Exploring non-invasive sampling of parasites by metabarcoding gastrointestinal nematodes in Madagascar frog species. Basic and Applied Herpetology. 2018;32:29-40. PubMed PMID: Zoorec:Zoor15503013639.

69. Avramenko RW, Redman EM, Lewis R, Yazwinski TA, Wasmuth JD, Gilleard JS. Exploring the Gastrointestinal "Nemabiome": Deep Amplicon Sequencing to Quantify the Species Composition of Parasitic Nematode Communities. PLoS One. 2015;10(12). doi: 10.1371/journal.pone.0143559. PubMed PMID: 26630572 PMCID: PMC4668017.

70. Poissant J, Gavriliuc S, Bellaw J, Redman EM, Avramenko RW, Robinson D, et al. A repeatable and quantitative DNA metabarcoding assay to characterize mixed strongyle infections in horses. Int J Parasitol. 2021;51(2-3):183-92. Epub 2020/11/27. doi: 10.1016/j.ijpara.2020.09.003. PubMed PMID: 33242465.

71. Wang C, Zhang T, Wang Y, Katz LA, Gao F, Song W. Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. Proc Biol Sci. 2017;284(1859). Epub 2017/07/28. doi: 10.1098/rspb.2017.0425. PubMed PMID: 28747472; PubMed Central PMCID: PMCPMC5543213.

72. Bourret V, Gutierrez Lopez R, Melo M, Loiseau C. Metabarcoding options to study eukaryotic endoparasites of birds. Ecol Evol. 2021;11(16):10821-33. Epub 2021/08/26. doi: 10.1002/ece3.7748. PubMed PMID: 34429884; PubMed Central PMCID: PMCPMC8366860.

73. Flaherty BR, Barratt J, Lane M, Talundzic E, Bradbury RS. Sensitive universal detection of blood parasites by selective pathogen-DNA enrichment and deep amplicon sequencing. Microbiome. 2021;9(1):1. Epub 2021/01/04. doi: 10.1186/s40168-020-00939-1. PubMed PMID: 33388088; PubMed Central PMCID: PMCPMC7778815.

74. Flaherty BR, Talundzic E, Barratt J, Kines KJ, Olsen C, Lane M, et al. Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in

blood via targeted amplicon deep sequencing. Microbiome. 2018;6(1):164. Epub 2018/09/19. doi: 10.1186/s40168-018-0540-2. PubMed PMID: 30223888; PubMed Central PMCID: PMCPMC6142370.

75. Gilbert JA, Jansson JK, Knight R. Earth Microbiome Project and Global Systems Biology. mSystems. 2018;3(3). Epub 2018/04/17. doi: 10.1128/mSystems.00217-17. PubMed PMID: 29657969; PubMed Central PMCID: PMCPMC5893859.

76. Belda E, Coulibaly B, Fofana A, Beavogui AH, Traore SF, Gohl DM, et al. Preferential suppression of *Anopheles gambiae* host sequences allows detection of the mosquito eukaryotic microbiome. Sci Rep. 2017;7(1):3241. Epub 2017/06/14. doi: 10.1038/s41598-017-03487-1. PubMed PMID: 28607435; PubMed Central PMCID: PMCPMC5468309.

77. Krogsgaard LR, Andersen LO, Johannesen TB, Engsbro AL, Stensvold CR, Nielsen HV, et al. Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome. Clin Transl Gastroenterol. 2018;9(6):161. Epub 2018/06/20. doi: 10.1038/s41424-018-0027-2. PubMed PMID: 29915224; PubMed Central PMCID: PMCPMC6006308.

Figures

Figure 1.

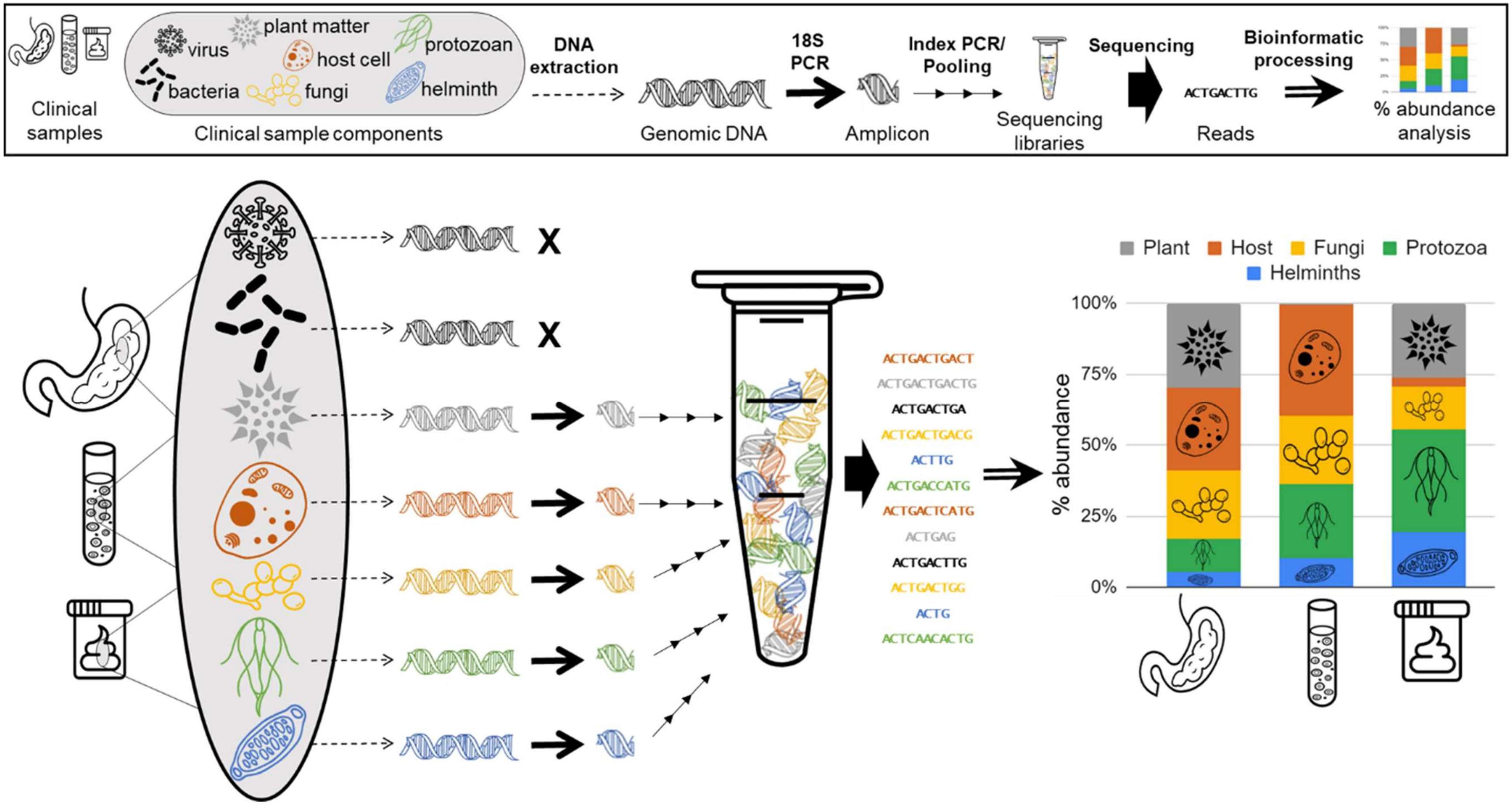


Figure 1. Schematic overview of 18S metabarcoding methods. Clinical samples such as stomach contents, blood, or feces are subjected to genomic DNA extraction. The resulting mixture contains DNA from all component organisms in the sample, which may include viruses, bacteria, plants, host, fungi, protozoa, and helminths. Extracted DNA is used as template in a PCR using 18S primers. Note that DNA from organisms that do not have 18S genes (*e.g*., viruses, bacteria, archaea) will not be bound by 18S primers and therefore will not result in a PCR product (denoted by an **X**). Resulting amplicons are processed to add indexes and adapters, pooled, and subjected to DNA sequencing. Resulting data consist of short nucleotide sequences ("reads") which are filtered for quality and assigned taxonomy using bioinformatic tools. Composition of samples is represented by percent abundance analysis showing the relative number of reads originating from organisms or groups of organisms.

Figure 2.

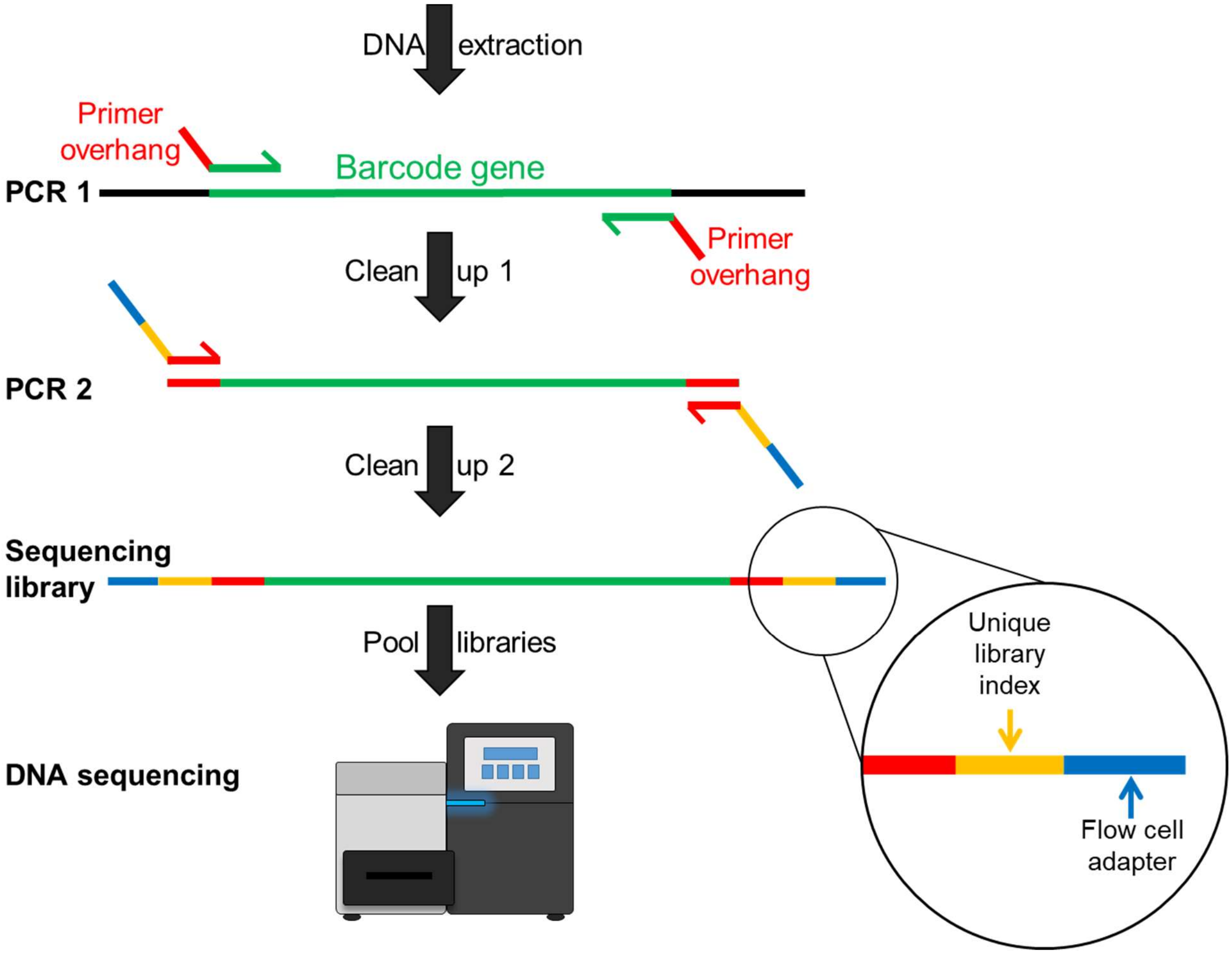


Figure 2. Schematic overview of sequencing library preparation using 2-step protocol. PCR primers used in PCR 1 ("Amplicon PCR") contain a locus-specific sequence targeting a barcode region (green) linked to overhang sequences (red). PCR products are purified and used as template for PCR 2 ("Index PCR") using primers complementary to the overhang sequences (red) and containing unique index sequences (yellow) and flow cell binding sequences (blue). Resulting products (*i.e.,* sequencing libraries) are purified, quantified, and pooled with additional libraries prior to DNA sequencing.

CHAPTER 2: A *Sarcina* bacterium linked to lethal disease in sanctuary chimpanzees in Sierra Leone

Leah A. Owens, Barbara Colitti, Ismail Hirji, Andrea Pizarro, Jenny E. Jaffe, Sophie Moittié, Kimberly A. Bishop-Lilly, Luis A. Estrella, Logan J. Voegtly, Jens H. Kuhn, Garret Suen, Courtney L. Deblois, Christopher D. Dunn, Carles Juan-Sallés, & Tony L. Goldberg. 2021 A *Sarcina* bacterium linked to lethal disease in sanctuary chimpanzees in Sierra Leone. *Nature Communications.* **12**, no. 1, 1-16. (doi: 10.1038/s41467-021-21012-x)

Abstract

Human and animal infections with bacteria of the genus *Sarcina* (family Clostridiaceae) are associated with gastric dilation and emphysematous gastritis. However, the potential roles of sarcinae as commensals or pathogens remain unclear. Here, we investigate a lethal disease of unknown etiology that affects sanctuary chimpanzees (*Pan troglodytes verus*) in Sierra Leone. The disease, which we have named "epizootic neurologic and gastroenteric syndrome" (ENGS), is characterized by neurologic and gastrointestinal signs and results in death of the animals, even after medical treatment. Using a case-control study design, we show that ENGS is strongly associated with *Sarcina* infection. The microorganism is distinct from *Sarcina ventriculi* and other known members of its genus, based on bacterial morphology and growth characteristics. Whole-genome sequencing confirms this distinction and reveals the presence of genetic features that may account for the unusual virulence of the bacterium. Therefore, we propose that this organism be considered the representative of a new species, named "*Candidatus* Sarcina troglodytae". Our results suggest that a heretofore unrecognized complex of related sarcinae likely exists, some of which may be highly virulent. However, the potential role of "*Ca*. S. troglodytae" in the etiology of ENGS, alone or in combination with other factors, remains a topic for future research.

Introduction

Emerging pathogens pose a substantial risk to animal and human health (1, 2). Pathogens can emerge due to the acquisition of virulence factors through genetic mutation and horizontal gene transfer (3-5) and due to ecological processes that alter their epidemiology/epizootiology and host range (6-8). These processes are accelerating due to factors such as agricultural intensification (9), demographic shifts (10), ecosystem perturbations (11), geophysical processes (12), and global environmental changes (13). Furthermore, improved diagnostic methods facilitate detection and characterization of new pathogens and the genetic features that contribute to their emergent phenotypes (14-16).

Emerging pathogens of non-human primates are especially salient examples of this phenomenon because of the high potential for such pathogens to infect humans, who are genetically-similar hosts (17, 18). For example, one of the main causative agents of human malaria, *Plasmodium falciparum*, once thought to have co-evolved with humans, actually arose from a recent zoonotic transmission from a western gorilla (*Gorilla gorilla* (Savage, 1847)) (19-21). Nowhere are such risks more evident than in zoological and sanctuary settings, where captive and semi-captive primates come into frequent close contact with people (22, 23). For example, contact with New World primates led to simian foamy virus transmission to primate workers in Brazil (24) and monkeypox virus transmission occurred in staff at a primate sanctuary following a monkeypox outbreak in sanctuary chimpanzees (*Pan troglodytes* (Blumenbach, 1775)) in Cameroon (25).

Since 2005, western chimpanzees (*Pan troglodytes verus* Schwarz, 1934; “chimpanzees” hereafter) in Sierra Leone’s Tacugama Chimpanzee Sanctuary (TCS, in Western Area Peninsula National Park) have suffered from a lethal disease of unknown etiology. Characteristic signs are neurologic (weakness, ataxia, seizures) and gastrointestinal (abdominal distension, anorexia, vomiting), resulting in death even after aggressive medical treatment by staff veterinarians. To date, a total of 56 individuals at this facility have died of this condition, which we have named "epizootic neurologic and gastroenteric syndrome" (ENGS), constituting a medical emergency in this population, which averages 93 chimpanzees at any given time. TCS is the largest repository of Sierra Leone's chimpanzee genetic diversity, a training site for conservationists throughout western Africa, an educational/ecotourism destination important for the local economy, and the only home for displaced or orphaned chimpanzees in the country.

Despite enormous efforts by veterinary staff and international collaborators, the etiology of ENGS has remained elusive. Encephalomyocarditis virus (EMCV; *Picornaviridae*: *Cardiovirus*) infection and toxicity from certain plants (*Dichapetalum toxicarium* (G. Don) Baill./*D. heudelotii* (Planch. ex Oliv.) Baill.) were both suspected but, after investigation, deemed unlikely to be causal. In such circumstances where infection has been suspected but known agents have not been identified, diagnostic approaches based on metagenomics have proven useful (26-28). We therefore undertook a case-control epizootiological investigation to identify potential pathogens of all major types (viruses, bacteria, and eukaryotes) using metagenomics and traditional methods in the TCS chimpanzees, to detect associations between particular microbes and ENGS.

Here we report the finding of a novel *Sarcina* genus bacterium in 13 of 19 ENGS cases but no controls. We also report the occurrence of gross and histopathological lesions in affected chimpanzees consistent with the most severe forms of *Sarcina* infection reported in humans and animals. By studying the morphology and growth characteristics of the new bacterium, and by sequencing the complete genome of an isolate, we identify features that distinguish it from all previously described members of its genus. In particular, we show that the new bacterium possesses genes encoding biochemical pathways potentially contributing to enhanced virulence, including an encoded urea degradation biochemical pathway, consistent with the clinical signs observed in chimpanzees. We conclude that the genus *Sarcina* likely contains an overlooked complex of species ranging from benign commensals to frank pathogens. In light of these findings, the importance of sarcinae in human and animal clinical disease should be re-evaluated.

Results

Epizootiology, clinical signs, and pathology

From 2005 - 2018, 56 resident chimpanzees of TCS died of ENGS. In 32 of 56 cases, affected individuals displayed signs including anorexia, neuromuscular weakness, ataxia, seizures, vomiting, and abdominal distension (Fig. 1; “Clinical signs” group). Signs persisted for a median of 6 days (range: 1-90 days) prior to recovery or death (Supplementary Table 1). In all recovered cases, clinical disease subsequently recurred and resulted in death. In the remaining 24 cases, individuals were discovered post-mortem with no premonitory signs noted by care staff or developed signs which progressed to death in 12 hours or less (Fig. 1; “Sudden death” group). Despite these disparate manifestations, all clinical presentations were associated and clearly recognizable as the same “mystery disease” (described as “unmistakable” by veterinarians). We therefore chose the term “syndrome” to reflect the heterogenous nature of the clinical presentations and the suspicion of a common etiopathogenesis. ENGS represented the highest cause of mortality in this population, affecting 33.7% of chimpanzees and accounting for 63.6% of deaths during this time period (Fig. 2a), with a case fatality rate of 100% and a seasonal distribution peaking in March (Fig. 2b). The etiological agent of ENGS does not appear to be transmitted directly, as there were few instances of cases clustering in time and space (Supplementary Table 1).

Frequent lesions included acute shock (congestion involving multiple organs), neutrophilic margination in the microcirculation, moderate to marked gastric dilation, pulmonary edema, acute aspiration of digestive contents, and acute hemorrhage in the thymus, pancreas, or both. In total, post-mortem evaluation documentation was available for 28 chimpanzees, but in only 17 of these was the gastrointestinal tract assessed. Of these 17 patients, 14 had gross evidence of acute gastric dilation, 1 did not show such evidence, and 2 were inconclusive (Fig. 1). One of the chimpanzees with acute gastric dilation also had massive hemorrhagic diathesis (Fig. 3a) and multiple gas-filled lesions in the cecal wall (emphysematous typhlocolitis; Fig. 3b). Microscopically, these gas-filled lesions were surrounded by infiltrates of macrophages, eosinophils, and multinucleate giant cells (Fig. 3c).

Samples

We selected 95 archived samples from 32 chimpanzees for analysis (Supplementary Table 2). These samples comprised 19 individuals (7 males and 12 females) that had died from ENGS (cases), representing a subset of the 56 total cases that occurred since the epizootic began in 2005 (Fig. 2), and 14 healthy individuals (7 males and 7 females) sampled during routine veterinary health checks or, in 2 instances, sampled post-mortem when cause of death was known and clearly unrelated to ENGS (*e.g.*, from trauma; controls). In one instance, matched samples were available from a healthy chimpanzee who subsequently became ill and died from ENGS (hence this individual was first a control and then a case). The chimpanzees in this study ranged in age from 5 to 27 years (median age = 12 years) and were sampled between 14 March 2013 and 11 July 2016, although not all cases within this time period were sampled or available for study. Among ENGS cases included, clinical signs were similar to those of cases that were not available for inclusion (Fig. 1), the most common of which were ataxia (n = 12), seizures (n = 12), vomiting (n = 10), and abdominal distention (n = 9).

Parasitology

Microscopic examinations of fecal samples were performed on site for 30 chimpanzees (17 ENGS cases and 13 controls) from 2005 - 2018 using standard direct and sedimentation methods (29), comprising 155 analyses with a median 5 analyses per chimpanzee. Nine records indicated no detectable parasites and 146 records indicated ≥1 parasite, with a median parasite richness of 2. Parasites identified included *Entamoeba* spp., *Troglodytella abrassarti*, *Balantidium coli, Ascaris* spp., *Enterobius* spp., *Strongyloides* spp., *Trichostrongylus* spp., *Trichuris* spp*., Taenia* spp*., Schistosoma* spp., and unspecified flagellated protozoa and hookworms (Supplementary Fig. 1), all of which are common in this population of chimpanzees and in apparently healthy animals in other captive and wild settings (29-31). Fortuitously, three ENGS cases had undergone fecal parasitological examinations immediately prior to or just after the time of death, and only representatives of these same typical/commensal organisms were identified: *Troglodytella abrassarti*, *Entamoeba hartmanni, Balantidium coli,* and *Enterobius* spp..

Eukaryotic metabarcoding for parasite identification using Earth Microbiome Project (EMP) protocols (32) and previously-published primers (33, 34) for 12 ENGS cases and 6 controls (24 samples; Supplementary Table 3) yielded a total of 2,955,014 raw reads (1,477,507 paired reads) with good overall sequencing quality (~21% of reads removed during quality filtering; Supplementary Table 4). Due to the pan-eukaryotic nature of the primers, and despite the use of a mammal-blocking primer, the majority of reads were identified as host (~83%; Supplementary Table 4). This identification was not surprising due to the sample types (host tissues) and because of previously published similar findings using the same primers and protocols (35). We processed data from all samples through all filtering steps, after which we excluded those samples that represented less than 0.5% of the total filtered data set, resulting in removal of 7 of 24 samples (Supplementary Table 4). From the remaining reads (range 681 to 31,131 per sample) we identified 7 operational taxonomic units (OTUs) representing 3 parasitic and 4 environmental eukaryotic organisms (Supplementary Fig. 2). No parasites thus identified were case-associated (*i.e.* found at statistically significantly different prevalence in case versus control groups using a Fisher's exact test (two-tailed); Supplementary Table 5).

Virology

Metagenomics for virus discovery conducted using previously published methods (36-38) on 12 ENGS cases and 6 controls (24 samples; Supplementary Table 3) generated a total of 151,206,140 reads (mean per sample 6,300,256, standard deviation 3,738,306; average length 161, standard deviation 27). After trimming and filtering on length and quality, 71.3% of sequences remained (mean per sample 4,495,233, standard deviation 2,792,589; average length 113.3, standard deviation 16.3) which were then assembled into 952 contiguous sequences (contigs hereafter; mean per sample 39.7, standard deviation 45.5; average length per contig 969.5, standard deviation 690.8) at an average sequence depth of 63.4 (standard deviation 172.4, minimum 3.5, maximum 851). Overall, 21.0% of reads assembled into contigs and 79.0% did not. Analyses of sequence data at the individual read level confirmed the results of the analysis of contigs (*i.e*. identified the same viruses) and did not identify any additional viruses. Eleven viruses were thus identified, each of which was identical to or very closely related to a known virus (Supplementary Table 6), and no viruses were case-associated (Supplementary Table 5). One control animal was infected with a rhinovirus C (*Picornaviridae*: *Enterovirus*) subsequent to an outbreak of respiratory illness in the population. This pathogen was previously documented as a cause of epizootic respiratory disease in wild chimpanzees (37), but the clinical features of this rhinovirus C infection (characterized by upper respiratory signs) are not consistent with ENGS.

Bacterial Metabarcoding

PCR amplification of the 16S rDNA V4 region was attempted on all 96 samples (19 cases and 14 controls; Supplementary Table 2) which yielded amplicons in 10 of 19 ENGS cases (35 total samples) but none of the controls (0 samples). In total, 1,131,561 raw sequences were generated for all 35 samples of which 787,263 were high quality after filtering in mothur. Good's coverage estimation was calculated for all samples and only those samples that had a value > 0.99 were retained for downstream analysis. As a result, 23 samples were considered for this analysis (Supplementary Table 3), which totaled 774,339 high quality sequences with an average of 33,667 +/- 17,122 standard deviation per sample. These sequences were binned at 97% similarity into 2,592 OTUs. The OTU counts were normalized to 5,900 sequences per sample, and these normalized sequences were used for all further analyses.

Analysis of these samples showed that 9 of 23 samples contained a large proportion of sequences (>5% of total reads and up to 97.4% in one sample) belonging to a single OTU belonging to an unknown member of the bacterial family *Clostridiaceae* (*Clostridia*: *Clostridiales*) (Fig. 4a). This OTU most closely resembled *Clostridium perfringens* in the

Greengenes database (39, 40). However, *C. perfringens* diagnostic PCR using published protocols (41-43) failed to yield amplicons in any instances, including in tissues found positive by 16S sequencing.

Re-examination of the representative sequence from this OTU against the National Center for Biotechnology Information's (NCBI's) GenBank (GenBank hereafter) non-redundant database excluding uncultured organisms identified a putative match (97.2% nucleotide identity) to *Clostridium* (*Sarcina*) *ventriculi* from feces of Japanese macaques (*Macaca fuscata* Blyth, 1875; GenBank accession numbers LC101491 and LC101492). Re-examination of all samples by including the *Clostridium* (*Sarcina*) *ventriculi* sequence from GenBank in the Greengenes database demonstrated this organism to be present in all 23 ENGS case samples (Supplementary Table 3). Notably, the organism was present not only in gastrointestinal contents but also in internal organs such as brain, liver, and spleen, sometimes at very high abundance (Fig. 4b). The nomenclature of the genus *Sarcina* is contested (and sometimes the genus name *Clostridium* is substituted) because *Sarcina* is phylogenetically situated within the "cluster I" group of *Clostridia* (Johnson and Francis, 1975) (44), considered the "true" *Clostridia,* although these organisms are polyphyletic. A proposal was made to change the name *Sarcina* to *Clostridium*, but was not approved because the name *Sarcina* predates the name *Clostridium* and therefore has priority (45).

Diagnostic PCR

Oligonucleotide primers specific to the 16S rDNA gene of the unknown organism were successfully developed. PCR with these primers yielded amplicons of the predicted length (289 base pairs [bp]) in 13 of 19 ENGS cases (68.4%) but 0 of 13 controls (Supplementary Fig. 3) which was statistically significant (odds ratio = 56.1; 95% CI 2.87–1097.2; Fisher's exact *P* = 0.0001, two-tailed; Supplementary Table 5). For one individual, blood samples were available both before and after clinical illness and death from ENGS; the pre-illness blood sample (collected in February 2016) was PCR-negative whereas the post-mortem blood sample (collected in July 2016) was PCR-positive. Sanger sequences of all amplicons were identical to each other and to the representative sequence generated by metabarcoding, except for one sample with a single nucleotide polymorphism (C→T transition) at position 51 of the diagnostic fragment. Sequences of the diagnostic fragment were also identical to published 16S rDNA sequences in GenBank for *S. ventriculi* (AF110272) and *Clostridium ventriculi* DSM286 (NR026146), with the exception of the one variant sequence (1 nucleotide mismatch to the aforementioned published sequences).

Bacterial isolation and characterization

We attempted to culture the bacterium using 44 combinations of cell preparations and culture conditions (Supplementary Table 7), 2 of which resulted in growth of colonies that resembled sarcinae. Specifically, wet mounts of colonies grown on egg yolk agar plates and *Sarcina ventriculi* growth medium (SVGM) plates revealed refractile, cuboid cells in packets, a morphology that is distinctive of members of the genus *Sarcina* (Fig. 5a, right panel). These colonies were derived from the liver of one individual (1 colony on egg yolk agar plates) and the brain of another (many colonies on SVGM plates; Supplementary Table 3). We confirmed the identity of every colony using diagnostic PCR and Sanger sequencing (see above).

We were repeatedly able to isolate the organism by plating brain tissue onto SVGM plates, but the organism ceased to remain viable after 2–3 passages and did not grow in any of the 7 liquid media tested (Supplementary Table 7). These results are consistent with previous studies reporting great difficulty in isolating and propagating sarcinae (46, 47). Furthermore, we were unable to recover live organisms after freezing colonies placed in 10% or 20% glycerol under various conditions. In contrast, we successfully grew the type strain *S. ventriculi* “Goodsir” (American Type Culture Collection [ATCC] 29068) under the same conditions with ease, including propagating the strain in solid and liquid media, freezing the strain in 10% glycerol, and subsequently recovering the bacterium. On SVGM media, our isolate (JB1) grew more slowly than *S. ventriculi* “Goodsir” (approximately 3–4 days until colonies were visible for JB1, versus 24 h for *S. ventriculi* “Goodsir”). The JB1 isolate displayed morphology (Fig. 5a), Gram’s staining characteristics (Fig. 5b), and methylene blue staining characteristics (Fig. 5c) similar to *S. ventriculi* “Goodsir”, as both were Gram-positive with a darkly staining outer layer. However, JB1 cells were statistically significantly larger than those of *S. ventriculi* “Goodsir” (mean diameters of 4.29 µm versus 2.83 µm, respectively, Mann-Whitney U $P = 0.0006$, two-tailed; Fig. 5d). The cellular diameter of JB1 falls within the published range for *S. maxima* (4 – 4.5 µm) (48), but methylene blue stain showed a cellulose-containing cell wall for isolate JB1 which is not characteristic of *S. maxima* (Fig. 5c). In addition, the flattened cellular morphology and large packet size of JB1 cells resemble *S. ventriculi* and not *S. maxima* (49-51).

Archived histologic preparations of tissues collected from ENGS cases during postmortem examination and stained with hematoxylin and eosin clearly revealed sarcinae, visible as packets of darkly staining basophilic cells in gastric contents of the chimpanzee with hemorrhagic diathesis, gastric dilation and emphysematous gastritis, and in the pulmonary alveoli of another chimpanzee (Supplementary Fig. 4a). A wet mount direct smear of homogenized brain tissue from the aforementioned ENGS case also demonstrated the presence of packets of sarcinae (Supplementary Fig. 4b).

16S rDNA phylogeny

Alignment of 16S rDNA sequences from the "cluster I" group of *Clostridia* (Johnson and Francis, 1975) (44), including the new organism (isolate JB1), yielded a final alignment length of 1,585 positions. A maximum likelihood phylogeny built from this alignment shows the new bacterium to represent a sister taxon to *S. ventriculi*, forming a clade with *S. maxima, Eubacterium tarantellae*, and *C. perfringens* (Fig. 6). Bacteria of 13 other recognized species pairs included in the analysis had a lower phylogenetic distance between them than the distance between the new bacterium and *S. ventriculi* (Supplementary Table 8), lending support to the designation of the new bacterium as a representative of a novel species. To reflect the discovery of this bacterium in chimpanzees (*Pan troglodytes* spp.), we designated it "*Candidatus* S. troglodytae". We propose the *Candidatus* designation in this instance because we were unable to generate a culture suitable for deposition in the requisite two publicly-accessible culture repositories in two different countries (52).

Whole-genome sequencing, assembly, and annotation

To generate sufficient material for whole-genome sequencing, we repeated bacterial isolation from the brain tissue described above using identical methods and allowed colonies to grow to a large size on SVGM plates. We then harvested a single, large colony, confirmed its identity using microscopy and PCR/sequencing, and extracted DNA from this colony (isolate JB2). We performed whole-genome sequencing using a hybrid approach of mate-pair and shotgun sequencing (Supplementary Table 9) followed by *de novo* genomic assembly using SPAdes (53) and *in silico* genome closure. The resulting full, high-quality "*Ca*. S. troglodytae" assembly (accessions CP051754 – CP051764) consists of a circular chromosome of 2,435,860 base pairs resolved into one single contig and 10 plasmids (totaling 205,993 base pairs, range: 4.6–78.9kb; Supplementary Table 10).

We annotated the genome with PATRIC (54) and confirmed that the new organism is closely related, but not identical, to *S. ventriculi* (98.5% nucleotide similarity; Supplementary Table 11) (55, 56). Total GC content in our organism was 27.6%, which is very similar to that of *S. ventriculi* (27.7%). The new organism shares 96.5% of its open reading frames (ORFs) with *S. ventriculi*, with notable differences in sugar pathways and capsule biosynthesis (Fig. 7). The genome of the JB2 strain contains DNA elements encoding metabolic pathways with the potential for formation of bacterial endospores in addition to anaerobic fermentation pathways, including alcohol fermentation, sulfur reduction, and nitrogen reduction. Interestingly, the "*Ca*. S. troglodytae" genome, but not *S. ventriculi*, possesses ORFs encoding for urea degradation enzymes, including the urease sub-units alpha, beta, and gamma, and ORFs whose products are predicted to be urease accessory proteins UreE, UreF, and UreG, which are needed for urease maturation. Ureases are nickel-containing enzymes, found in a variety of bacteria, which catalyze the breakdown of urea (a ubiquitous metabolic byproduct in most animals) to ammonia and carbon dioxide (57). Bacterial and fungal ureases play a key role in gastrointestinal tract colonization and in chronic human diseases such as gastritis and peptic ulcers (58). Additionally, in the extrachromosomal sequences, we found a 34 kb plasmidial prophage containing ORFs for an ATP-binding cassette (ABC) transporter. ABC transporters use ATP to move specific substrates across a cellular membrane and can function either as an importer (*e.g.*, for uptake of nutrients) or an exporter (*e.g.*, to efflux toxic molecules, including xenobiotic compounds such as drugs) (59). ABC importers have been associated with increased bacterial survival during colonization of hosts (60) and exporters with bacterial drug-resistance (61), both of which may enhance pathogenesis of an organism.

Sarcinae are not known to produce toxins (62). However, because of the unusual neurologic disease associated with ENGS, we scanned the genome sequence of "*Ca*. S. troglodytae" for ORFs with sequence homology to known virulence genes using ShortBRED (63) and a customized version of the Virulence Factor Database (VFDB) (64), but found no evidence of such genes. Using the Comprehensive Antibiotic Resistance Database (CARD) (65), we identified two antibiotic resistance ORFs, *OXA-241* on the chromosome and *salA* on plasmid 1, which confer resistance to carbapenems (*OXA-241*) and lincosamides and streptogramins (*salA*) (66, 67).

Summary description of the provisional species

"*Candidatus* Sarcina troglodytae" is a proposed member of the established genus *Sarcina*, most closely-related to *S. ventriculi,* as determined by full-length 16S rDNA phylogenetic analysis. It is an uncultivated, Gram-positive coccus with a tetrad structure and slightly flattened cell morphology and may be identified using the PCR primers: TacuSarc_Diag_F: 5′-TGAAAGGCATCTTTTAACAATCAAAG-3′ and TacuSarc_Diag_R: 5′-TACCGTCATTATCGTCCCTAAA-3′ or the full genome sequence (accessions CP051754 – CP051764). We isolated "*Ca.* S. troglodytae" in an anerobic environment and at mesophilic temperature (37° C) but were unable to maintain a viable culture for deposition in at least two publicly accessible culture collections, hence the *Candidatus* status. Samples described here are derived from the brain, liver, and lung tissues of sanctuary western chimpanzees (*Pan troglodytes verus*) diagnosed with epizootic neurologic and gastroenteric syndrome (ENGS).

Discussion

The genus *Sarcina* within the *Clostridiaceae* is poorly studied in comparison to the highly-studied toxigenic clostridia. In 1842, Goodsir described the type species, *S. ventriculi*, in the stomach contents of a human patient with recurrent vomiting (68). Subsequent studies have provided evidence that bacteria morphologically consistent with *S. ventriculi* cause abdominal pain, nausea, anorexia, vomiting, hematemesis, dysphagia, diarrhea, and generalized weakness in people (69), with esophagitis (70) and duodenitis (71) noted surgically or as a post-mortem finding (72). Morphologically indistinguishable bacteria assumed to be *S. ventriculi* have also been associated with abomasal bloat in young pre-ruminant animals (73-75), characterized by sudden onset of anorexia, abdominal discomfort, lethargy, dehydration, and shock culminating in high lethality (75–100%) despite treatment (76). Gastric dilation in monogastric animals (horses, dogs, and cats) has also been linked to putative *S. ventriculi* infection (77, 78).

Our results demonstrate a statistically significant association between a new bacterium, "*Ca*. S. troglodytae", and ENGS, a protracted lethal epizootic syndrome in sanctuary chimpanzees in Sierra Leone. We designed a case-control epizootiological study using a case definition that encompassed the range of clinical presentations associated with the syndrome, which included both sudden death and gastrointestinal and neurologic signs prior to death. This case definition will likely become more refined as the syndrome is studied further. By applying metabarcoding and metagenomics, we did not find differences between case and control groups with respect to infection with any parasite or virus and therefore deemed these types of organisms unlikely to be causes of ENGS. However, bacterial metabarcoding and a subsequent PCR revealed infection with "*Ca*. S. troglodytae" in 68.4% of ENGS cases but no controls. In one instance, a chimpanzee was PCR-negative for "*Ca*. S. troglodytae" when healthy but subsequently became PCR-positive after succumbing to ENGS.

Sarcinae are notoriously difficult to culture, particularly from non-environmental sources such as animal tissues (46, 47). Despite being studied since the 1800s, sarcinae have been isolated successfully from only a handful of animal or human sources (75, 79, 80) (see Supplementary Table 12 for review). Prior to this study, only one photomicrograph of unfixed sarcinae cells in their native morphology was published (81). Although we were able to isolate JB1, it did not survive repeated passages or freezing, distinguishing it from its closest relative, *S. ventriculi*, as does its larger cell size in culture and slower growth. Flattened cell morphology and a cellulose-containing cell wall distinguish "*Ca*. S. troglodytae" from *S. maxima* (51), its next closest relative, despite overlapping cell size. Phylogenetic analysis based on 16S rDNA demonstrates the difference between "*Ca*. S. troglodytae" and *S. ventriculi* to be greater than the difference between bacteria of 13 other recognized species pairs within the clostridial rDNA group I (82). Whole-genome sequencing and genetic characterization revealed 69 ORFs that were not found in the genome of *S. ventriculi* "Goodsir." For these reasons, we propose that this organism be considered the representative of a new species within the genus *Sarcina*.

Bacteria within the family *Clostridiaceae* include organisms linked to life-threatening diseases, as well as benign commensals and environmental bacteria (62). Patterns of virulence/toxigenicity do not correspond to phylogeny, and pathogenicity cannot be predicted based on 16S rDNA sequence grouping alone (83). Whole genome sequencing of "*Ca*. S.

troglodytae" revealed no toxin ORFs similar to those present in toxigenic clostridia (84). The pathogenic effects of "*Ca*. S. troglodytae" on chimpanzees may therefore be caused by mechanisms other than toxicity. For example sarcinae have an unusual yeast-like metabolism (85) that is active over a wide pH range (86), allowing bacteria to produce carbon dioxide gas and ethanol prolifically, both of which can cause disease in the gastrointestinal tract and the central nervous system (87, 88).

We also found that the genome sequence of "*Ca*. S. troglodytae" contains ORFs encoding for a predicted urease. Although urease expression is associated with normal microbial flora in some instances, ureases are better known as a key virulence factors in pathogenic bacteria such as *Clostridium perfringens*, *Helicobacter pylori*, and *Klebsiella pneumoniae* (89) and are associated with diseases including ammonia encephalopathy, hepatic encephalopathy, hepatic coma, and gastroduodenal infections (57). Because, in other bacteria, ureases have established roles in infection and persistence in the host (90), stimulation of host inflammatory reactions (91), cytotoxic effects on host cells (92), and damage to extracellular matrix (93) and tight junctions (94), the presence of a urease biochemical pathway in "*Ca*. S. troglodytae" could help explain the bacterium's pathogenesis and dissemination outside of the gastrointestinal tract. For example, urease activity in the yeast *Cryptococcus neoformans* is responsible for central nervous system invasion; unlike the wild-type organism, mutants lacking this enzyme do not disseminate to the brain and cause meningoencephalitis (95). Moreover, the major product of urea degradation, ammonia, could enhance the ability of "*Ca*. S. troglodytae" to cause neurologic signs, because ammonia is highly neurocytotoxic *in vivo* (96).

In some cases of *Sarcina* infection in humans, symptoms are preceded by evidence of delayed gastric emptying (69, 70, 97-99). With ENGS, however, affected chimpanzees appeared healthy prior to the onset of signs. It is therefore noteworthy that several studies have shown colonization of *Sarcina* and lesions in the absence of delayed gastric emptying (71, 100, 101). For example, a recent publication concerning a lethal case of human emphysematous gastritis highlights several similarities to ENGS, including lack of gastroparesis, afebrile and normotensive presentation, gastrointestinal and neurologic signs, and rapid death (102). The occurrence of acute gastric dilation and emphysematous lesions in the digestive tract of one ENGS case included in our study recalls cases of *Sarcina* infection in humans and other animal species, which include acute gastric dilation and emphysematous gastritis. A more consistent gross and histopathologic evaluation of affected chimpanzees in the future may reveal a higher proportion of affected chimpanzees because acute gastric dilation and emphysematous gastrointestinal lesions may be misinterpreted as autolysis or overlooked grossly. For example, a chimpanzee in this population who died of ENGS subsequent to the analyses presented here clearly showed emphysematous lesions throughout the gastrointestinal tract. Although each of the clinical characteristics of ENGS (abdominal distention, nausea, vomiting, anorexia, diarrhea, and neurologic deficits) are common to diverse diseases, they are all consistent with emphysematous gastroenteritis, as is acute lethality, which likely results from irreversible hemodynamic instability and resulting systemic shock in emphysematous gastritis cases (103).

Notably, we found "*Ca*. S. troglodytae" not only in the gastrointestinal tracts of affected individuals, but also in internal organs, including the brain. For DNA extraction and culture, all tissues were maintained on dry ice, carefully sectioned using sterile technique, and subsampled from the innermost area while still frozen, leading us to conclude that our findings are not likely due to environmental contamination and instead reflect true infection. In human cases of *S. ventriculi* infection, there is precedent for bacteremia, likely originating from gastrointestinal translocation (104, 105), but to our knowledge presence of viable sarcinae in the central nervous system has not been previously reported. Central nervous system colonization may therefore be an overlooked clinical feature of severe *Sarcina* infection, or "*Ca*. S. troglodytae" may be a particularly virulent bacterium within the genus *Sarcina*. We advocate that prior and future human and animal cases of severe disease associated with sarcinae, particularly those cases without clear predisposing factors, be revisited, as they could represent other heretofore unrecognized presentations of infection with *Sarcina* bacteria. We also speculate that many documented cases of infection that were assumed to be caused by *S. ventriculi* based on morphology alone may actually have been caused by infection with taxonomically distinct sarcinae. If so, the genus *Sarcina* may contain a complex of morphologically cryptic species varying from benign environmental bacteria to lethal pathogens.

Many questions regarding ENGS and "*Ca*. S. troglodytae" remain unexplained. For example, epizootiologically, ENGS incidence peaks in March each year. As with other disease-associated clostridia, sarcinae form environmentally stable spores (106) and may be ubiquitous in soil (79, 99, 107), but environmental factors may contribute to germination of spores and overgrowth. Seasonal changes known to be important in chimpanzees, such as habitat use, diet (including exposure to plant or arthropod toxins), and physiological condition (108), may increase infection risk at certain times of the year. Alternatively, "*Ca*. S. troglodytae" may be re-introduced seasonally, for instance by migratory animals (109), perhaps with seasonal weather patterns facilitating its establishment (110). The potential role of "*Ca*. S. troglodytae" in the etiology of ENGS, alone or in combination with other factors, remains a topic for future research.

Due to lack of infrastructure at TCS and limitations on sample shipments to the U.S., our analysis could only include samples obtained between 2013 and 2016, even though ENGS was first noted in 2005. Ideally we would have obtained the same tissues, especially from the gastrointestinal tract, from all cases, in addition to complete medical records and post-mortem examination notes including gross and histological findings. Unfortunately, this was not possible due to the resource-limited setting and resulting opportunistic sampling. Fortunately, the veterinary staff at TCS have begun standardizing sample and record collection for future cases.

Practical and ethical considerations preclude collecting invasive samples from sanctuary chimpanzees, so control samples were limited to serum and feces collected at annual health checks and samples collected post-mortem from individuals determined to have died from causes other that ENGS (*e.g.*, accidental death), which are very rare at TCS. Furthermore, the evidence presented here supports an association between "*Ca.* S. troglodytae" and ENGS which, due to ethical considerations, cannot be further investigated by experimentation on chimpanzees. Likewise, infection trials in an animal model (*e.g.*, laboratory mice) would require a pure culture and, as of yet, we are unable to maintain a culture of this new bacterium, hence the *Candidatus* designation (52).

We also note that clinically-similar cases have not been reported in other captive or wild populations of chimpanzees or other primates. Moreover, despite over 10 years of illness among the TCS chimpanzees, human cases have not been reported, even among personnel with close daily contact with affected individuals. The genetic and physiological similarities between humans and chimpanzees are often cited as predisposing them to cross-species pathogen exchange (17, 111). It is therefore surprising that no human disease similar to ENGS has been reported to date. Should "*Ca*. S. troglodytae" indeed affect chimpanzees but not humans, it would represent a rare example of such a pathogen (112). However, we cannot rule out physiological stressors, diet-related factors, environmental conditions, or other pathogens as predisposing factors that differ between humans and chimpanzees. For example, in the case of *C. perfringens*-associated enteritis in humans, changes in gastric and intestinal pH, altered nutritional status, and concurrent infection, particularly with intestinal viruses and parasites, can drastically alter clinical outcomes (113).

To our knowledge, only 44 cases of *Sarcina* infection in humans have been reported in the peer-reviewed literature since the beginning of the 1900s (Supplementary Table 12), and currently no standard treatment for such infections is available. Of published cases treated with at least one specifically-mentioned antibiotic (19 of 44), the most common regimen was a combination of oral ciprofloxacin and metronidazole (11 of 19) with a proton-pump inhibitor (8 of 11) or antacid (2 of 11). With the exception of one case involving other complications, treatment was successful when follow-up was noted (9 of 10). Four published cases detail dosages of the antibiotics, all in adult males, and most dosages (3 of 4) were identical: 250 mg metronidazole three times daily and 250 mg ciprofloxacin twice daily for a course of 7 days. Recently, *S. ventriculi* cultured from human blood was shown to be susceptible to other antibiotics including penicillin (minimal inhibitory concentration [MIC] = 0.25 mg/l), amoxicillin (MIC = 0.50 mg/l), amoxicillin-clavulanic acid, piperacillin-tazobactam, imipenem, clindamycin, levofloxacin, rifampicin, vancomycin, and linezolid (104).

Treatment of emphysematous gastritis is similarly unstandardized and includes hemodynamic stabilization with intravenous fluids, broad spectrum intravenous antibiotics effective against Gram-negative and anaerobic bacteria, including meropenem (114, 115), cefuroxime and metronidazole (88), nafcillin and cefoxitin (116), and surgery in some cases, but is associated with 60% lethality (103). That we found evidence of two antibiotic resistance genes in our "*Ca*. S. troglodytae" isolate, a chromosomal *OXA-241*-like gene involved in carbapenem resistance and a plasmid-associated *salA*-like gene linked to lincosamide/streptogramin resistance (67), is noteworthy, as these findings may influence ENGS treatment decisions. Overgrowth of sarcinae in the stomach appears to predispose patients to clinical disease (72); therefore, probiotics, particularly those containing acidophilic organisms, may prove useful for the treatment or prevention of "*Ca*. S. troglodytae" infections in chimpanzees. For example, probiotics have proven useful for the prevention of *C. difficile*-related disease in humans (117, 118). Finally, autogenous vaccines have proven useful for the prevention of *C. perfringens*-related disease in animals (119, 120). Such an approach could prove useful for the prevention of ENGS if *in vitro* growth conditions for "*Ca*. S. troglodytae" can be determined.

Since 2011, case studies and reviews concerning *S. ventriculi* and human disease have increased in the medical literature from 0 articles from 1900–2000, to 2 articles from 2000–2010 and 33 articles from 2011 – November 2019 (Supplementary Table 12). Increased recent attention to *Sarcina* despite establishment of the genus in 1842 may be coincidental. Alternatively, it may indicate a nascent trend of bacterial emergence (121, 122). The physiological and environmental drivers of *Sarcina* acquisition and subsequent disease progression merit greater attention than they have heretofore received, as does the genetic diversity of the genus. In 34 of 44 published cases, diagnosis of *Sarcina* infection was based on morphology and/or Gram staining alone with no other diagnostics for confirmation (Supplementary Table 12). Cases of clinical disease associated with *Sarcina* infections should be re-evaluated in light of the possibility that the bacteria identified may represent a complex of cryptic species and strains, some of which are benign but others of which may be highly virulent.

Methods

Ethics Statement

Tacugama Chimpanzee Sanctuary (TCS) located in Western Area National Park, Sierra Leone, is a non-governmental organization that operates under the purview and with the permission of the Ministry of Agriculture, Forestry, and Food Security. All animals originated from Sierra Leone and were confiscated or handed over to TCS under the authority of the Ministry. TCS does not remove any animals from the wild but works to rescue chimpanzees that have been removed from the wild illegally. The care and sampling of resident chimpanzees is officially sanctioned by the Government of Sierra Leone, and samples were shipped to the USA with the official permission of the Government of Sierra Leone under Convention on International Trade in Endangered Species of Wild Fauna and Flora permit number 17US19807C/9.

The presented study was retrospective, did not involve collection of any samples solely for the purpose of this research, and utilized surplus samples collected by TCS veterinarians during routine veterinary procedures and post-mortem examination, which are standard at the sanctuary for any fatality, in compliance with the “Pan African Sanctuary Alliance Primate Veterinary Healthcare Manual” (123) and the policies of TCS.

Clinical data and samples

We obtained clinical data from veterinary records for chimpanzees who had died of all causes from 2005 through 2018. These data were compiled by year and by month to make epizootic curves. We then used these data to select samples from the TCS freezer archive according to a case-control study design. Due to resource limitations, samples were collected opportunistically (as opposed to systematically) and archived samples were only available from a subset of cases that occurred from 2013 to 2016. Samples had been collected by staff veterinarians during routine health checks or during post-mortem examination, and samples were fresh-frozen (at -20 °C or -80 °C) upon collection in whirl packs or test tubes (Supplementary Table 2) and stored long-term at -80 °C. Samples were shipped frozen on dry ice to the United States, stored at -80 °C upon arrival, and kept frozen through processing. To obtain sub-samples of solid tissues and avoid contamination from the external surfaces of organs, we cut frozen tissues with a sterile razor blade and extracted tissue plugs from the newly-exposed area with a sterile 6-mm biopsy punch.

Parasitology

Microscopy for parasite identification was performed at TCS from 2005 to 2018 following standard veterinary protocols (29). Briefly, freshly voided fecal samples were collected from individuals and macroscopic features were noted. A direct smear was then made by mixing fecal material with saline and observing the mixture under a light microscope at 100X and 400X total magnification, with an additional formalin-ether (10% formalin and ethyl acetate) sedimentation performed as warranted. Slides were read by trained and experienced staff veterinarians. Data on the occurrence of parasites thus identified were compiled from 155 such analyses conducted from 2005 through 2018, representing 17 ENGS-affected chimpanzees (cases) and 13 apparently healthy chimpanzees (controls).

Molecular parasitology using metabarcoding was performed with methods modified from the EMP (32). DNA was extracted from tissue samples (blood, plasma, serum, lung, and brain) using the DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) according to manufacturer’s instructions and eluted in 50 µl of buffer AE (10 mM Tris-HCl, 0.5 mM ethylenediaminetetraacetic acid). Tissue samples were considered appropriate for this analysis based on published literature showing that infections, including with eukaryotes, can be detected in such samples, even when the tissues analyzed are not the anatomic sites of infection (124, 125). Primers were used to amplify the V9 region of the 18S rDNA gene and were based on published pan-eukaryotic sequences (126, 127). These sequences were modified, replacing the individual barcodes with overhang sequences compatible with the Nextera system (Illumina, San Diego, CA, USA)(33-35). The primers used were F: 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACACACCGCCCGTC-3′ and R: 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGATCCTTCTGCAGGTTCACCTAC-3′ (IDT, Newark, NJ, USA). To reduce host signal, we used the EMP mammal blocking primer: 5′-GCCCGTCGCTACTACCGATTGGII IIITTAGTGAGGCCCT-[C3 Spacer]-3′ (IDT) and performed PCR according to EMP protocols.

PCR products were purified using the DNA Clean and Concentrator Kit (Zymo Research, Irvine, CA, USA) and eluted in 25 µl of provided elution buffer. From the 25 µl, 5 µl was then used as a template in a 25-µl PCR mix with the Nextera XT Index Kit v2 (Illumina) and limited-cycle PCR for indexing using an annealing temperature of 55 °C with 12 cycles. Products were separated on a 1.5% agarose gel and visualized to confirm band lengths of approximately 260 bp. Amplicons were then excised from gels and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research) and eluted in 20 µl of water. Products were quantified using a Qubit fluorometer (Thermo-Fisher Scientific Inc., Waltham, MA, USA). Libraries were sequenced on a MiSeq instrument using paired-end 300 x 300 cycle chemistry (Illumina).

Raw reads were processed with the QIIME v.1.9.1 pipeline (128). Forward and reverse reads were assembled into paired contigs using the command multiple_join_paired_ends.py and quality filtered using the command multiple_split_libraries_fastq.py with default parameters, except for setting the Phred threshold to 30 or higher (-q 29) and minimum length to 100 bp (-l 100). Chimeras were identified with Usearch v6.1 (129) and removed. Reads were then assigned to OTUs using the QIIME protocol for open reference OTU picking with the command pick_open_reference_otus.py and the default UCLUST tool (129), and taxonomy was assigned to OTUs using default settings with the command assign_taxonomy.py against the SILVA database version 132 (130). Still-undetermined OTUs were assigned using BLAST within QIIME (-m blast) against the full GenBank database (131) and non-target sequences were then removed by filtering. We processed all data from through all filtering steps after which we removed those samples that represented less than 0.5% of the total filtered data set from further analyses.

Virology

Samples were homogenized by bead beating (for solid tissues), clarified by centrifugation, treated with nucleases and processed for metagenomic virus discovery as previously described (36-38). Briefly, viral RNA was isolated using a QIAamp MinElute virus spin kit (Qiagen), omitting carrier RNA. Extracted nucleic acids were then converted to double-stranded cDNA using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Carlsbad, CA, USA) and random hexamers and purified using Ampure XP beads (Beckman Coulter, Brea, CA, USA). Approximately 1 ng of DNA was prepared as a library for pair-ended sequencing on a MiSeq instrument (MiSeq Reagent kit v3, 150 cycle) using the Nextera XT DNA Library Prep Kit (Illumina). Sequence data were analyzed using CLC Genomics Workbench version 11.0 (Qiagen). In brief, we trimmed low-quality bases (Phred quality score <30), discarded short reads (<75 bp), and subjected the remaining reads to *de novo* assembly using the CLC assembler with automatic word and bubble size selection and a minimum contig length of 500. We then analyzed contigs for nucleotide- (blastn) and protein-level (blastx) similarity to known viruses in GenBank. For blastx we applied the BLASTX algorithm with the BLOSUM62 matrix to sequences translated into all 6 frames. We also analyzed all sequence data at the individual read level by mapping reads to viruses in the GenBank database using the CLC mapping tool at low stringency (length fraction of 0.5, similarity fraction of 0.6).

Bacterial Metabarcoding

Genomic DNA was extracted from solid tissue and blood samples using a DNeasy Blood and Tissue Kit and from fecal, vomit, and stomach content samples using a DNeasy PowerSoil DNA Isolation Kit (Qiagen) according to manufacturer's instructions. The V4 region of the bacterial 16S rRNA gene was amplified using universal primers (132). PCR was performed as previously described (133). Briefly, reactions were carried out in 25 µl volumes containing 10 µM of each primer, 12.5 µl 2x HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA), 6.5 µl water, and 25 ng template DNA with the following cycling conditions: 95 °C for 3 min; 30 cycles of 95 °C for 30 sec, 55 °C for 30 sec, 72 °C for 30 sec; and 72 °C for 5 min. PCR products were then elecrophoresed on 1% low melt agarose gels (National Diagnostics, Atlanta, GA, USA), excised, purified using a ZR-96 Zymoclean Gel DNA Recovery Kit (Zymo Research), and quantified using a Qubit® Fluorometer (Thermo Fisher Scientific). Equimolar amounts of the barcoded V4 amplicons were pooled and sequenced using a MiSeq 2×250 bp v2 kit (Illumina) using custom sequencing primers with 10% PhiX control DNA.

All sequences were demultiplexed on the Illumina MiSeq and were processed and analyzed using mothur v.1.42.0 according to standard methods (132). Briefly, poor quality sequences were removed after paired end sequences were combined into contigs. Sequences were aligned against the SILVA 16S rRNA gene reference alignment database to screen for alignment to the correct region. Preclustering was performed to reduce error and chimeras were detected and removed using UCHIME (134). The SILVA database was used to classify bacterial sequences while sequences classifying to mitochondria, cyanobacteria, Eukarya, *Archaea*, or Fungi were removed along with singletons to streamline analysis.

Diagnostic PCR

PCR primers were designed to the V2-V3 region of the "*Ca*. S. troglodytae" 16S rRNA gene: TacuSarc_Diag_F: 5′-TGAAAGGCATCTTTTAACAATCAAAG-3′ (T_m = 52.8 °C) and TacuSarc_Diag_R: 5′-TACCGTCATTATCGTCCCTAAA-3′ (T_m = 53 °C) (IDT). PCR reactions were carried out in 25 µl volumes containing 0.2 µM of each primer, 12.5 µl 2x HotStar Master

Mix (Qiagen), 10 µl water, and 25 ng template DNA on a C-1000 thermocycler (BioRad, Hercules, CA, USA) with the following cycling conditions: 95 °C for 15 min; 29 cycles of 94 °C for 30 sec, 48 °C for 30 sec, 72 °C for 30 sec; and 72 °C for 10 min. PCR products (289 bp expected length) were then electrophoresed on 1.5% low-melt agarose gels with ethidium bromide and 1 Kb Plus DNA length standards (BioRad), visualized under UV light, and photographed using a GelDoc XR imager (BioRad). Amplicons were then excised and purified as described above and Sanger sequenced on ABI 3730xl DNA Analyzers (Applied Biosystems, Foster City, CA, USA) at the University of Wisconsin-Madison Biotechnology Center.

Bacterial isolation and characterization

Liquid samples were pipetted directly onto sterile agar plates and solid tissues (deep cut sections collected with a 6-mm biopsy punch to avoid external contamination) were placed in sterile petri dishes and minced with 2 sterile blades until homogenized. 200 µl of pre-reduced thioglycollate medium (Hardy Diagnostics, Santa Maria, CA, USA) was added and the mixture was streaked by inoculating loop onto a 100 mm x 15 mm plate, placed immediately into an Anaerogen Compact anaerobic pouch (Oxoid Limited, Hampshire, UK), sealed, and incubated at 37 °C. For liquid growth media, a sterile 18-gauge needle with 1-ml syringe was used to inoculate stoppered tubes that were incubated anaerobically at 37 °C. Cultures were screened by PCR, and cells were directly visualized and grown at least 10 days before they were deemed negative for growth of the bacterium of interest.

For comparison, the type strain *S. ventriculi* “Goodsir” (ATCC 19633 or ATCC 29068) was obtained from the American Type Culture Collection (American Type Culture Collection, Manassas, VA, USA) and grown according to ATCC guidelines.

Bacterial imaging

Live bacterial cells diluted in sterile water or phosphate buffered saline were placed on glass slides, examined with light microscopy, and imaged immediately. For Gram staining, heat-fixed slides were flooded with crystal violet solution for 1 min, rinsed with water, flooded with iodine solution for 1 min, rinsed with water, flooded with decolorizer solution for 1–5 sec, rinsed with water, counterstained with 5 drops of safranin solution for 30 sec, rinsed with water, and air dried. For methylene blue staining, heat-fixed slides were flooded with 1% aqueous solution of methylene blue for 1 min at room temperature, then washed with distilled water and then air dried. All slides were visualized and photographed at 400X on a Panthera U microscope with a Moticam 5.0 camera (Motic, British Columbia, Canada). For cell size measurements, strains JB1 and ATCC 29068 "Goodsir" were plated and grown on SVGM plates under identical conditions for 72 h. Bacterial cells from 7 distinct colonies were harvested, blinded to the investigator, and examined as follows: live cells in phosphate buffered saline from 5 non-overlapping visual fields were captured as above and single-cell diameters were quantified using the circle (3-point) measurement tool in the Images Plus software suite (Motic). Cellular diameters were compared using a Mann-Whitney U test (two-tailed).

Tissues for histopathology were collected during post-mortem examination by staff veterinarians and immediately fixed in 4% paraformaldehyde at least overnight, then later dehydrated in alcohol, embedded in paraffin wax, cut into 6-µm sections, stained with hematoxylin and eosin, visualized under a light microscope, and photographed. For direct visualization of brain tissue, a 6-mm biopsy punch (Integra LifeSciences, Plainsboro, NJ, USA) was taken from the interior of the cerebrum, minced with sterile blades, smeared onto a clean glass slide, and immediately imaged as described above.

16S rDNA phylogeny

The full 16S rDNA sequence from "*Ca*. S. troglodytae" isolate JB1 (1,508 bp) was queried against the NCBI 16S ribosomal RNA sequence (*Bacteria* and *Archaea*) database using megablast (131) with default parameters, and the top 50 results as of 17 July 2019 were downloaded from NCBI's RefSeq or GenBank (all e-values were 0). For taxonomic completeness, 73 sequences comprising "cluster I" *Clostridia* (135) as previously published (136) were also retrieved from GenBank, and duplicates were removed. The type organism from the closest known clade (*Hathewaya histolytica*, located in "cluster II" of the clostridia as defined by Collins et al. 1994 (135)) was included as an outgroup (136). The resulting 98 sequences, plus the new "*Ca*. S. troglodytae" sequence, were aligned using MUSCLE3.8.31 (137) (final alignment length 1,585 positions). To quantify nucleotide-level distances among sequences, a pairwise distance matrix was calculated using MEGA7 (138) with pairwise deletion and 1,000 bootstrap replicates to estimate standard errors. The phylogenetic position of "*Ca*. S. troglodytae" was then inferred with PhyML v1.8.1 (139) using the General Time Reversible (GTR) substitution model as determined by Smart Model Selection (140), and 1,000 bootstrapped data sets were used to estimate statistical confidences of clades.

Whole-genome sequencing, assembly, and annotation

A large, single colony of cells morphologically consistent with "*Ca*. S. troglodytae" (Fig. 5a) which tested positive by diagnostic PCR was grown for 7 days on an SVGM plate at 37 °C in an anaerobic pouch. The entire colony ("isolate JB2") was transferred into a sterile 1.5-ml tube and genomic DNA was extracted using the Wizard Genomic DNA purification kit (Promega) according to the manufacturer's instructions. Mate-pair sequencing libraries were constructed using 1 µg of resulting DNA and the Nextera gel-free protocol and quantified using NEBNext qPCR (New England Biolabs, Ipswich, MA, USA). All libraries were loaded in equimolar amounts, multiplexed, and sequenced on an Illumina MiSeq using the 2 x 300 bp V3 chemistry.

The mate-pair reads were processed with Nxtrim (141), then processed with bbduk (142) using Phred = 20, length > 50 base pairs, and subsampled to 100x coverage with bbnorm (143). We used bowtie2 (144) to remove sample-to-sample bleed through. All reads were assembled with SPAdes v3.10.1 (53) using -careful and -mp arguments and manually closed with Bandage (145), CLC Workbench 12.0 and EDGE Bioinformatics (146). The PATRIC server (147) with default settings was used for all ORF annotations, CLC Genomics Workbench 12.0 was used for variant analysis, Mauve (148) was used for genome comparisons, ISFinder (149) was used for insertion sequence identification, the Resistance Gene Identifier (RGI v4.2.2) from the Comprehensive Antibiotic Resistance Database (CARD v3.0.0) (65) was used for antibiotic resistance gene identification, a customized database from the Virulence Factor Database (VFD) (64) was used for virulence factor identification with ShortBRED (63), and ANI (Average Nucleotide Identity) Calculator (55) was used to generate the average nucleotide identities.

Data Availability

Sequence data that support the findings of this study have been deposited in the National Center for Biotechnology Information (NCBI) GenBank database with the accession codes CP051754 – CP051764 (Bacterial genome assembly) and MT350347 – MT350357 (Viral replicase genes) and in the NCBI Sequence Read Archive with the accession codes SRR11551921 – SRR115511922 (Bacterial genome sequencing reads) and BioProject accession code PRJNA648419 (16S and 18S metabarcoding reads).

References

1. Brower JL. The threat and response to infectious diseases (Revised). Microb Ecol. 2018;76(1):19-36.

2. Cunningham AA, Daszak P, Wood JLN. One Health, emerging infectious diseases and wildlife: two decades of progress? Philosophical Transactions of the Royal Society B-Biological Sciences. 2017;372(1725).

3. Lakhundi S, Zhang K. Methicillin-resistant *Staphylococcus aureus*: molecular characterization, evolution, and epidemiology. Clin Microbiol Rev. 2018;31(4).

4. Gomez-Simmonds A, Uhlemann AC. Clinical implications of genomic adaptation and evolution of carbapenem-resistant *Klebsiella pneumoniae*. J Infect Dis. 2017;215(suppl_1):S18-S27.

5. Harrison LH, Simonsen V, Waldman EA. Emergence and disappearance of a virulent clone of *Haemophilus influenzae* biogroup aegyptius, cause of Brazilian purpuric fever. Clin Microbiol Rev. 2008;21(4):594-605.

6. Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, et al. Pathways to zoonotic spillover. Nature Reviews Microbiology. 2017;15(8):502-10.

7. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016;14(8):523-34.

8. Rogalski MA, Gowler CD, Shaw CL, Hufbauer RA, Duffy MA. Human drivers of ecological and evolutionary dynamics in emerging and disappearing infectious disease systems. Philos Trans R Soc Lond B Biol Sci. 2017;372(1712).

9. Rohrl JR, Barrett CB, Civitello DJ, Craft ME, Delius B, DeLeo GA, et al. Emerging human infectious diseases and the links to global food production. Nature Sustainability. 2019;2(6):445-56.

10. Johnson PT, de Roode JC, Fenton A. Why infectious disease research needs community ecology. Science. 2015;349(6252):1259504.

11. Hernandez H, Martinez LR. Relationship of environmental disturbances and the infectious potential of fungi. Microbiology. 2018;164(3):233-41.

12. McMahon BJ, Morand S, Gray JS. Ecosystem change and zoonoses in the Anthropocene. Zoonoses Public Health. 2018;65(7):755-65.

13. Hoberg EP, Brooks DR. Evolution in action: climate change, biodiversity dynamics and emerging infectious disease. Philos Trans R Soc Lond B Biol Sci. 2015;370(1665).

14. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341-55.

15. Radoshevich L, Cossart P. *Listeria monocytogenes*: towards a complete picture of its physiology and pathogenesis. Nat Rev Microbiol. 2018;16(1):32-46.

16. Hoffmann C, Zimmermann F, Biek R, Kuehl H, Nowak K, Mundry R, et al. Persistent anthrax as a major driver of wildlife mortality in a tropical rainforest. Nature. 2017;548(7665):82-6.

17. Devaux CA, Mediannikov O, Medkour H, Raoult D. Infectious disease risk across the growing human-non human primate interface: a review of the evidence. Front Public Health. 2019;7:305.

18. Sharp PM, Rayner JC, Hahn BH. Evolution. Great apes and zoonoses. Science. 2013;340(6130):284-6.

19. Plenderleith LJ, Liu W, Learn GH, Loy DE, Speede S, Sanz CM, et al. Ancient introgression between two ape malaria parasite species. Genome Biol Evol. 2019;11(11):3269-74.

20. Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, et al. Out of Africa: origins and evolution of the human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. Int J Parasitol. 2017;47(2-3):87-97.

21. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7:11078.

22. Dunay E, Apakupakul K, Leard S, Palmer JL, Deem SL. Pathogen transmission from humans to great apes is a growing threat to primate conservation. Ecohealth. 2018;15(1):148-62.

23. Muehlenbein MP. Primates on display: potential disease consequences beyond bushmeat. Am J Phys Anthropol. 2017;162 Suppl 63:32-43.

24. Muniz CP, Cavalcante LTF, Jia H, Zheng H, Tang S, Augusto AM, et al. Zoonotic infection of Brazilian primate workers with New World simian foamy virus. PLoS One. 2017;12(9):e0184502.

25. Guagliardo SAJ, Monroe B, Moundjoa C, Athanase A, Okpu G, Burgado J, et al. Asymptomatic orthopoxvirus circulation in humans in the wake of a monkeypox outbreak among chimpanzees in Cameroon. Am J Trop Med Hyg. 2019.

26. Murkey JA, Chew KW, Carlson M, Shannon CL, Sirohi D, Sample HA, et al. Hepatitis E virus-associated meningoencephalitis in a lung transplant recipient diagnosed by clinical metagenomic sequencing. Open Forum Infect Dis. 2017;4(3):ofx121.

27. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med. 2014;370(25):2408-17.

28. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. N Engl J Med. 2019;380(24):2327-40.

29. Modrý DP, B. Petrželková, K. Hasegawa, H. Parasites of apes an atlas of coproscopic diagnostics: Edition Chimaira; 2018. 198 p.

30. Howells ME, Pruetz J, Gillespie TR. Patterns of gastro-intestinal parasites and commensals as an index of population and ecosystem health: the case of sympatric western chimpanzees (*Pan troglodytes verus*) and guinea baboons (*Papio hamadryas papio*) at Fongoli, Senegal. Am J Primatol. 2011;73(2):173-9.

31. McLennan MR, Hasegawa H, Bardi M, Huffman MA. Gastrointestinal parasite infections and self-medication in wild chimpanzees surviving in degraded forest fragments within an agricultural landscape mosaic in Uganda. PLoS One. 2017;12(7):e0180431.

32. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457-+.

33. Lappan R, Classon C, Kumar S, Singh OP, de Almeida RV, Chakravarty J, et al. Meta-taxonomic analysis of prokaryotic and eukaryotic gut flora in stool samples from visceral leishmaniasis cases and endemic controls in Bihar State India. PLoS Negl Trop Dis. 2019;13(9):e0007444.

34. Waraniak JM, Marsh TL, Scribner KT. 18S rRNA metabarcoding diet analysis of a predatory fish community across seasonal changes in prey availability. Ecol Evol. 2019;9(3):1410-30.

35. Belda E, Coulibaly B, Fofana A, Beavogui AH, Traore SF, Gohl DM, et al. Preferential suppression of *Anopheles gambiae* host sequences allows detection of the mosquito eukaryotic microbiome. Sci Rep. 2017;7(1):3241.

36. Goldberg TL, Bennett AJ, Kityo R, Kuhn JH, Chapman CA. Kanyawara virus: a novel rhabdovirus infecting newly discovered nycteribiid bat flies infesting previously unknown pteropodid bats in Uganda. Sci Rep. 2017;7(1):5287.

37. Scully EJ, Basnet S, Wrangham RW, Muller MN, Otali E, Hyeroba D, et al. Lethal respiratory disease associated with human Rhinovirus C in wild chimpanzees, Uganda, 2013. Emerg Infect Dis. 2018;24(2):267-74.

38. Toohey-Kurth K, Sibley SD, Goldberg TL. Metagenomic assessment of adventitious viruses in commercial bovine sera. Biologicals. 2017;47:64-8.

39. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72.

40. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610-8.

41. Rinttilä T, Kassinen A, Malinen E, Krogius L, Palva A. Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. J Appl Microbiol. 2004;97(6):1166-77.

42. Wang RF, Cao WW, Franklin W, Campbell W, Cerniglia CE. A 16S rDNA-based PCR method for rapid and specific detection of *Clostridium perfringens* in food. Mol Cell Probes. 1994;8(2):131-7.

43. Wise MG, Siragusa GR. Quantitative detection of *Clostridium perfringens* in the broiler fowl gastrointestinal tract by real-time PCR. Appl Environ Microbiol. 2005;71(7):3911-6.

44. Johnson JL, Francis BS. Taxonomy of the *Clostridia*: ribosomal ribonucleic acid homologies among the species. J Gen Microbiol. 1975;88(2):229-44.

45. Tindall BJ. Priority of the genus name *Clostridium Prazmowski* 1880 (Approved Lists 1980) vs *Sarcina Goodsir* 1842 (Approved Lists 1980) and the creation of the illegitimate combinations *Clostridium maximum* (Lindner 1888) Lawson and Rainey 2016 and *Clostridium ventriculi* (Goodsir 1842) Lawson and Rainey 2016 that may not be used. Int J Syst Evol Microbiol. 2016;66(11):4890-4.

46. Edwards GT, Woodger NG, Barlow AM, Bell SJ, Harwood DG, Otter A, et al. *Sarcina*-like bacteria associated with bloat in young lambs and calves. Vet Rec. 2008;163(13):391-3.

47. Panciera RJ, Boileau MJ, Step DL. Tympany, acidosis, and mural emphysema of the stomach in calves: report of cases and experimental induction. J Vet Diagn Invest. 2007;19(4):392-5.

48. Canale-Parola E. Biology of the sugar-fermenting *Sarcinae*. Bacteriol Rev. 1970;34(1):82-97.

49. Canale-Parola E, Mandel M, Kupper DG. The classification of sarcinae. Arch Mikrobiol. 1967;58(1):30-4.

50. Canale-Parola E, Wolfe RS. Synthesis of cellulose by *Sarcina ventriculi*. Biochim Biophys Acta. 1964;82:403-5.

51. Holt SC, Canale-Parola E. Fine structure of *Sarcina maxima* and *Sarcina ventriculi*. J Bacteriol. 1967;93(1):399-410.

52. International code of nomenclature of prokaryotes. Int J Syst Evol Microbiol. 2019;69(1A):S1-S111.

53. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.

54. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017;45(D1):D535-D42.

55. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints; 2016. Report No.: 2167-9843.

56. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57(Pt 1):81-91.

57. Burne RA, Chen YY. Bacterial ureases in infectious diseases. Microbes Infect. 2000;2(5):533-42.

58. Konieczna I, Zarnowiec P, Kwinkowski M, Kolesinska B, Fraczyk J, Kaminski Z, et al. Bacterial urease and its role in long-lasting human diseases. Curr Protein Pept Sci. 2012;13(8):789-806.

59. Rees DC, Johnson E, Lewinson O. ABC transporters: the power to change. Nat Rev Mol Cell Biol. 2009;10(3):218-27.

60. Tanaka KJ, Song S, Mason K, Pinkett HW. Selective substrate uptake: the role of ATP-binding cassette (ABC) importers in pathogenesis. Biochim Biophys Acta Biomembr. 2018;1860(4):868-77.

61. Lubelski J, Konings WN, Driessen AJ. Distribution and physiology of ABC-type transporters contributing to multidrug resistance in bacteria. Microbiol Mol Biol Rev. 2007;71(3):463-76.

62. Hatheway CL. Toxigenic clostridia. Clin Microbiol Rev. 1990;3(1):66-98.

63. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-specificity targeted functional profiling in microbial communities with ShortBRED. PLoS Comput Biol. 2015;11(12):e1004557.

64. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res. 2016;44(D1):D694-7.

65. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017;45(D1):D566-D73.

66. Higgins PG, Zander E, Seifert H. Identification of a novel insertion sequence element associated with carbapenem resistance and the development of fluoroquinolone resistance in *Acinetobacter radioresistens*. J Antimicrob Chemother. 2013;68(3):720-2.

67. Hot C, Berthet N, Chesneau O. Characterization of sal(A), a novel gene responsible for lincosamide and streptogramin A resistance in *Staphylococcus sciuri*. Antimicrob Agents Chemother. 2014;58(6):3335-41.

68. Goodsir J, Wilson G. History of a case in which a fluid periodically ejected from the stomach contained vagetable organisms of an undescribed form. Edinburgh Medical and Surgical Journal. 1842;51(151):430-43.

69. Laass MW, Pargac N, Fischer R, Bernhardt H, Knoke M, Henker J. Emphysematous gastritis caused by *Sarcina ventriculi*. Gastrointest Endosc. 2010;72(5):1101-3.

70. de Meij TGJ, van Wijk MP, Mookhoek A, Budding AE. Ulcerative gastritis and esophagitis in two children with *Sarcina ventriculi* infection. Front Med (Lausanne). 2017;4:145.

71. Canan O, Ozkale M, Kayaseicuk F. Duodenitis caused by *Sarcina ventriculi* in a case with Celiac disease and selective IgA deficiency. Cukurova Medical Journal. 2017;42(4):766-8.

72. Gaspar BL. The significance of *Sarcina* in routine surgical pathology practice. APMIS. 2016;124(6):436-43.

73. DeBey BM, Blanchard PC, Durfee PT. Abomasal bloat associated with *Sarcina*-like bacteria in goat kids. J Am Vet Med Assoc. 1996;209(8):1468-9.

74. Van Kruiningen HJ, Nyaoke CA, Sidor IF, Fabis JJ, Hinckley LS, Lindell KA. Clostridial abomasal disease in Connecticut dairy calves. Can Vet J. 2009;50(8):857-60.

75. Vatn S, Tranulis MA, Hofshagen M. *Sarcina* -like bacteria, *Clostridium fallax* and *Clostridium sordellii* in lambs with abomasal bloat, haemorrhage and ulcers. J Comp Pathol. 2000;122(2-3):193-200.

76. Marshall TS. Abomasal ulceration and tympany of calves. Vet Clin North Am Food Anim Pract. 2009;25(1):209-20, viii.

77. Im JY, Sokol S, Duhamel GE. Gastric dilatation associated with gastric colonization with *Sarcina*-like bacteria in a cat with chronic enteritis. J Am Anim Hosp Assoc. 2017;53(6):321-5.

78. Vatn S, Gunnes G, Nybø K, Juul HM. Possible involvement of *Sarcina ventriculi* in canine and equine acute gastric dilatation. Acta Vet Scand. 2000;41(3):333-7.

79. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature. 2016;533(7604):543-+.

80. Tolentino LE, Kallichanda N, Javier B, Yoshimori R, French SW. A case report of gastric perforation and peritonitis associated with opportunistic infection by *Sarcina ventriculi*. Laboratory Medicine. 2003;34(7):535-7.

81. Vatn S. Abomasal disease in lambs : aetiological and predisposing factors associated with bloat, haemorrhage and ulcers. Oslo: S. Vatn; 1999. 35 s p.

82. Lawson PA, Llop-Perez P, Hutson RA, Hippe H, Collins MD. Towards a phylogeny of the clostridia based on 16S rRNA sequences. FEMS Microbiol Lett. 1993;113(1):87-92.

83. Popoff MR, Bouvet P. Genetic characteristics of toxigenic *Clostridia* and toxin gene evolution. Toxicon. 2013;75:63-89.

84. Carter GP, Cheung JK, Larcombe S, Lyras D. Regulation of toxin production in the pathogenic clostridia. Molecular Microbiology. 2014;91(2):221-31.

85. Stephenson MP, Dawes EA. Pyruvic acid and formic acid metabolism in *Sarcina ventriculi* and the role of ferredoxin. J Gen Microbiol. 1971;69(3):331-43.

86. Goodwin S, Zeikus JG. Physiological adaptations of anaerobic bacteria to low pH: metabolic control of proton motive force in *Sarcina ventriculi*. J Bacteriol. 1987;169(5):2150-7.

87. Adinoff B, Bone GH, Linnoila M. Acute ethanol poisoning and the ethanol withdrawal syndrome. Med Toxicol Adverse Drug Exp. 1988;3(3):172-96.

88. Al-Jundi W, Shebl A. Emphysematous gastritis: case report and literature review. Int J Surg. 2008;6(6):e63-6.

89. Mora D, Arioli S. Microbial urease in health and disease. PLoS Pathog. 2014;10(12):e1004472.

90. Schoep TD, Fulurija A, Good F, Lu W, Himbeck RP, Schwan C, et al. Surface properties of *Helicobacter pylori* urease complex are essential for persistence. PLoS One. 2010;5(11):e15042.

91. Scopel-Guerra A, Olivera-Severo D, Staniscuaski F, Uberti AF, Callai-Silva N, Jaeger N, et al. The impact of *Helicobacter pylori* urease upon platelets and consequent contributions to inflammation. Front Microbiol. 2017;8:2447.

92. Dunn BE, Phadnis SH. Structure, function and localization of *Helicobacter pylori* urease. Yale J Biol Med. 1998;71(2):63-73.

93. Dubreuil JD, Giudice GD, Rappuoli R. *Helicobacter pylori* interactions with host serum and extracellular matrix proteins: potential role in the infectious process. Microbiol Mol Biol Rev. 2002;66(4):617-29, table of contents.

94. Wroblewski LE, Shen L, Ogden S, Romero-Gallo J, Lapierre LA, Israel DA, et al. *Helicobacter pylori* dysregulation of gastric epithelial tight junctions by urease-mediated myosin II activation. Gastroenterology. 2009;136(1):236-46.

95. Olszewski MA, Noverr MC, Chen GH, Toews GB, Cox GM, Perfect JR, et al. Urease expression by *Cryptococcus neoformans* promotes microvascular sequestration, thereby enhancing central nervous system invasion. Am J Pathol. 2004;164(5):1761-71.

96. Norenberg MD, Rama Rao KV, Jayakumar AR. Signaling factors in the mechanism of ammonia neurotoxicity. Metab Brain Dis. 2009;24(1):103-17.

97. DiMaio MA PW, Longacre TA. Gastric *Sarcina* organisms in a patient with cystic fibrosis. Human Pathology: Case Reports. 2014;1(3):45-8.

98. Gulati R, Khalid S, Tafoya MA, McCarthy D. Nausea and vomiting in a diabetic patient with delayed gastric emptying: do not delay diagnosis. Dig Dis Sci. 2019;64(3):681-4.

99. Smit J. The biology of the fermenting *Sarcinae*. Journal of Pathology and Bacteriology. 1933;36:455-68.

100. Alvin M, Al Jalbout N. Emphysematous gastritis secondary to *Sarcina ventriculi*. BMJ Case Rep. 2018;2018.

101. Ratuapli SK, Lam-Himlin DM, Heigh RI. *Sarcina ventriculi* of the stomach: a case report. World J Gastroenterol. 2013;19(14):2282-5.

102. Singh K. Emphysematous gastritis associated with *Sarcina ventriculi*. Case Rep Gastroenterol. 2019;13(1):207-13.

103. van Mook WN, van der Geest S, Goessens ML, Schoon EJ, Ramsay G. Gas within the wall of the stomach due to emphysematous gastritis: case report and review. Eur J Gastroenterol Hepatol. 2002;14(10):1155-60.

104. Bortolotti P, Kipnis E, Faure E, Faure K, Wacrenier A, Fauquembergue M, et al. Clostridium ventriculi bacteremia following acute colonic pseudo-obstruction: A case report. Anaerobe. 2019;59:32-4.

105. Tuuminen T, Suomala P, Vuorinen S. *Sarcina ventriculi* in blood: the first documented report since 1872. BMC Infect Dis. 2013;13:169.

106. Lowe SE, Pankratz HS, Zeikus JG. Influence of pH extremes on sporulation and ultrastructure of *Sarcina ventriculi*. J Bacteriol. 1989;171(7):3775-81.

107. Rings DM. Clostridial disease associated with neurologic signs: tetanus, botulism, and enterotoxemia. Vet Clin North Am Food Anim Pract. 2004;20(2):379-91, vii-viii.

108. Chancellor RL, Rundus AS, Nyandwi S. Chimpanzee seed dispersal in a montane forest fragment in Rwanda. Am J Primatol. 2017;79(3):1-8.

109. Bandelj P, Trilar T, Blagus R, Ocepek M, Rousseau J, Weese JS, et al. Prevalence and molecular characterization of *Clostridium difficile* isolated from European Barn Swallows (*Hirundo rustica*) during migration. BMC Vet Res. 2014;10:40.

110. Lipp EK, Farrah SA, Rose JB. Assessment and impact of microbial fecal pollution and human enteric pathogens in a coastal community. Mar Pollut Bull. 2001;42(4):286-93.

111. Calvignac-Spencer S, Leendertz SAJ, Gillespie TR, Leendertz FH. Wild great apes as sentinels and sources of infectious disease. Clinical Microbiology and Infection. 2012;18(6):521-7.

112. Sakuma R, Takeuchi H. SIV replication in human cells. Front Microbiol. 2012;3:162.

113. Allaart JG, van Asten AJ, Grone A. Predisposing factors and prevention of *Clostridium perfringens*-associated enteritis. Comp Immunol Microbiol Infect Dis. 2013;36(5):449-64.

114. Huang H, Wu S, Wang M, Zhang Y, Fang H, Palmgren AC, et al. *Clostridium difficile* infections in a Shanghai hospital: antimicrobial resistance, toxin profiles and ribotypes. Int J Antimicrob Agents. 2009;33(4):339-42.

115. Loi TH, See JY, Diddapur RK, Issac JR. Emphysematous gastritis: a case report and a review of literature. Ann Acad Med Singapore. 2007;36(1):72-3.

116. Moosvi AR, Saravolatz LD, Wong DH, Simms SM. Emphysematous gastritis: case report and review. Rev Infect Dis. 1990;12(5):848-55.

117. Goldenberg JZ, Yap C, Lytvyn L, Lo CK, Beardsley J, Mertz D, et al. Probiotics for the prevention of *Clostridium difficile*-associated diarrhea in adults and children. Cochrane Database Syst Rev. 2017;12:CD006095.

118. Johnston BC, Ma SS, Goldenberg JZ, Thorlund K, Vandvik PO, Loeb M, et al. Probiotics for the prevention of *Clostridium difficile*-associated diarrhea: a systematic review and meta-analysis. Ann Intern Med. 2012;157(12):878-88.

119. Gohari IM, Arroyo L, Macinnes JI, Timoney JF, Parreira VR, Prescott JF. Characterization of *Clostridium perfringens* in the feces of adult horses and foals with acute enterocolitis. Can J Vet Res. 2014;78(1):1-7.

120. Springer S, Finzel J, Florian V, Schoepe H, Woitow G, Selbitz HJ. [Occurrence and control of the *Clostridium perfringens* type A associated diarrhea of the suckling pigs with special consideration of the immunoprophylaxis]. Tierarztl Prax Ausg G Grosstiere Nutztiere. 2012;40(6):375-82.

121. Bushnell G, Mitrani-Gold F, Mundy LM. Emergence of New Delhi metallo-beta-lactamase type 1-producing enterobacteriaceae and non-enterobacteriaceae: global case detection and bacterial surveillance. Int J Infect Dis. 2013;17(5):e325-33.

122. Shurin PA, Marchant CD, Kim CH, Van Hare GF, Johnson CE, Tutihasi MA, et al. Emergence of beta-lactamase-producing strains of *Branhamella catarrhalis* as important agents of acute otitis media. Pediatr Infect Dis. 1983;2(1):34-8.

123. Unwin S, Cress D, Colin C, Bailey W, Boardman W. Primate veterinary health manual. Portland, OR: Pan African Sanctuary Alliance (PASA); 2009.

124. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, et al. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol. 2019;4(4):663-74.

125. Hong DK, Blauwkamp TA, Kertesz M, Bercovici S, Truong C, Banaei N. Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. Diagn Microbiol Infect Dis. 2018;92(3):210-3.

126. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One. 2009;4(7):e6372.

127. Stoeck T, Bass D, Nebel M, Christen R, Jones MD, Breiner HW, et al. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol. 2010;19 Suppl 1:21-31.

128. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-6.

129. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1.

130. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6.

131. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24(16):1757-64.

132. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013;79(17):5112-20.

133. Dennis PG, Seymour J, Kumbun K, Tyson GW. Diverse populations of lake water bacteria exhibit chemotaxis towards inorganic nutrients. ISME J. 2013;7(8):1661-4.

134. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27(16):2194-200.

135. Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia P, et al. The phylogeny of the genus *Clostridium:* proposal of five new genera and eleven new species combinations. Int J Syst Bacteriol. 1994;44(4):812-26.

136. Lawson PA, Rainey FA. Proposal to restrict the genus *Clostridium* Prazmowski to *Clostridium butyricum* and related species. Int J Syst Evol Microbiol. 2016;66(2):1009-16.

137. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.

138. Kumar R, Mishra BK, Lahiri T, Kumar G, Kumar N, Gupta R, et al. PCV: an alignment free method for finding homologous nucleotide sequences and its application in phylogenetic study. Interdiscip Sci. 2017;9(2):173-83.

139. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21.

140. Lefort V, Longueville JE, Gascuel O. SMS: smart model selection in PhyML. Mol Biol Evol. 2017;34(9):2422-4.

141. O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 2015;31(12):2035-7.

142. Joshi NA FJ. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files Version 1.33 ed2011.

143. Hornick L. BBNorm Guide: The Regents of the University of California; 2019 [Available from: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbnorm-guide/.

144. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.

145. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31(20):3350-2.

146. Li PE, Lo CC, Anderson JJ, Davenport KW, Bishop-Lilly KA, Xu Y, et al. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res. 2017;45(1):67-80.

147. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.

148. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.

149. Siguier P, Pérochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34(Database issue):D32-6.

150. Crowther JS. *Sarcina ventriculi* in human faeces. J Med Microbiol. 1971;4(3):343-50.

151. Edwards AN, McBride SM. Isolating and purifying *Clostridium difficile* spores. Methods Mol Biol. 2016;1476:117-28.

152. Lam-Himlin D, Tsiatis AC, Montgomery E, Pai RK, Brown JA, Razavi M, et al. *Sarcina* organisms in the gastrointestinal tract: a clinicopathologic and molecular study. Am J Surg Pathol. 2011;35(11):1700-5.

153. Sauter JL, Nayar SK, Anders PD, D'Amico M, Butnor KJ, Wilcox RL. Co-existence of *Sarcina* organisms and *Helicobacter pylori* gastritis/duodenitis in pediatric siblings. J Clin Anat Pathol (JCAP). 2013;1(1).

154. Louis GB, Singh P, Vaiphei K. *Sarcina* infection. BMJ Case Rep. 2014;2014.

155. Kumar M, Bhagat P, Bal A, Lal S. Co-infection of *Sarcina* and *Giardia* in a child. Oxf Med Case Reports. 2014;2014(7):118-9.

156. Karakus E, Kirsaclioglu CT. Coincidence of celiac disease with *Sarcina* infection. Turk J Gastroenterol. 2014;25 Suppl 1:318.

157. DiMaio MA, Park WG, Longacre TA. Gastric *Sarcina* organisms in a patient with cystic fibrosis. Human Pathology: Case Reports. 2014;1(3):45-8.

158. Bhagat P, Gupta N, Kumar M, Radotra BD, Sinha SK. A rare association of *Sarcina* with gastric adenocarcinoma diagnosed on fine-needle aspiration. J Cytol. 2015;32(1):50-2.

159. Berry AC, Mann S, Nakshabendi R, Kanar O, Cruz L. Gastric *Sarcina ventriculi*: incidental or pathologic? Ann Gastroenterol. 2015;28(4):495.

160. Carrigan S, Grin A, Al-Haddad S, Iakovlev V, Streutker C, Moore T, et al. Emphysematous oesophagitis associated with *Sarcina* organisms in a patient receiving anti-inflammatory therapy. Histopathology. 2015;67(2):270-2.

161. Chougule A, Muthu V, Bal A, Rudramurthy SM, Dhooria S, Das A, et al. Pulmonary gangrene due to *Rhizopus* spp., *Staphylococcus aureus*, *Klebsiella pneumoniae* and probable *Sarcina* organisms. Mycopathologia. 2015;180(1-2):131-6.

162. Sopha SC, Manejwala A, Boutros CN. *Sarcina*, a new threat in the bariatric era. Hum Pathol. 2015;46(9):1405-7.

163. Medlicott SAC. *Sarcina ventricularis* complicating a patient status post vertical banded gastroplasty, a case. Journal of Gastroenterology and Hepatology Research. 2015;4(2):1481-4.

164. Al Rasheed MR, Senseng CG. *Sarcina ventriculi* : review of the literature. Arch Pathol Lab Med. 2016;140(12):1441-5.

165. Bommannan K, Gaspar BL, Sachdeva MU. Pathogenic *Sarcina* in urine. BMJ Case Rep. 2016;2016.

166. Mironova M, Gobara N, Pennell CP, Sherwinter DA, Cimic A. *Sarcina ventriculi:* Aa case report of gastric perforation in 85-year-old male with history of colon cancer. Journal of Case Reports and Images in Pathology. 2017;3:20-3.

167. de Meij TG, van Wijk MP, Mookhoek A, Budding AE. Ulcerative Gastritis and esophagitis in two Children with Sarcina ventriculi Infection. Frontiers in medicine. 2017;4:145.

168. Behzadi J, Modi RM, Goyal K, Chen W, Pfeil S. *Sarcina ventriculi* as an unknown culprit for esophageal stricturing. ACG Case Rep J. 2017;4:e118.

169. Rajasekar S, Onteddu N, Gupta A. A rare case of emphysematous gastritis-*Sarcina ventriculi:* 1919. Am J Gastroenterol. 2018;113:S1091.

170. Liu L, Gopal P. *Sarcina ventriculi* in a patient with slipped gastric band and gastric distention. Clin Gastroenterol Hepatol. 2018;16(4):A25-A6.

171. Elvert JL, El Atrouni W, Schuetz AN. Photo Quiz: A bacterium better known by surgical pathologists than by clinical microbiologists. J Clin Microbiol. 2018;56(12).

172. Aggarwal S, Tyagi R, Selhi PK, Garg A, Sood A, Sood N. Coinfection of *Sarcina ventriculi* and *Candida* in a patient of gastric outlet obstruction: an overloaded pyloric antrum. Diagn Cytopathol. 2018;46(10):876-8.

173. Shetty NU, O'Connell J, Oshilaja OO, Patil DT, Procop GW, Sturgis CD. First documented case of *Sarcina* in esophageal brushing cytology. Diagn Cytopathol. 2018;46(10):886-7.

174. Bortolotti P, Kipnis E, Faure E, Faure K, Wacrenier A, Fauquembergue M, et al. *Clostridium ventriculi* bacteremia following acute colonic pseudo-obstruction: a case report. Anaerobe. 2019;59:32-4.

175. Singh H, Weber MA, Low J, Krishnan U. *Sarcina* in an adolescent with repaired esophageal atresia: a pathogen or a benign commensal? J Pediatr Gastroenterol Nutr. 2019;69(2):e57.

176. Propst R, Denham L, Deisch JK, Kalra T, Zaheer S, Silva K, et al. *Sarcina* organisms in the upper gastrointestinal tract: a report of 3 cases with varying presentations. Int J Surg Pathol. 2019:1066896919873715.

177. Dey B, Raphael V, Banik A, Khonglah Y. *Sarcina* in sputum cytology in a patient of pulmonary tuberculosis. J Cytol. 2019;36(4):219-21.

178. Zare SY, Kubik MJ, Savides TJ, Hasteh F, Hosseini M. A rare case of *Sarcina ventriculi* diagnosed on fine-needle aspiration. Diagn Cytopathol. 2019;47(10):1079-81.

179. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. J Infect Dis. 2020;221(Suppl 3):S292-S307.

180. Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, et al. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. Frontiers in Microbiology. 2014;5.

181. Mann AE, Mazel F, Lemay MA, Morien E, Billy V, Kowalewski M, et al. Biodiversity of protists and nematodes in the wild nonhuman primate gut. Isme J. 2020;14(2):609-22.

182. Jarman SN, McInnes JC, Faux C, Polanowski AM, Marthick J, Deagle BE, et al. Adelie Penguin Population Diet Monitoring by Analysis of Food DNA in Scats. Plos One. 2013;8(12).

183. Bhadury P, Austen MC. Barcoding marine nematodes: an improved set of nematode 18S rRNA primers to overcome eukaryotic co-interference. Hydrobiologia. 2010;641(1):245-51.

184. Avramenko RW, Bras A, Redman EM, Woodbury MR, Wagner B, Shury T, et al. High species diversity of trichostrongyle parasite communities within and between Western Canadian commercial and conservation bison herds revealed by nemabiome metabarcoding. Parasites & Vectors. 2018;11.

185. Avramenko RW, Redman EM, Lewis R, Bichuette MA, Palmeira BM, Yazwinski TA, et al. The use of nemabiome metabarcoding to explore gastro-intestinal nematode species diversity and anthelmintic treatment effectiveness in beef calves. International Journal for Parasitology. 2017;47(13):893-902.

186. Avramenko RW, Redman EM, Lewis R, Yazwinski TA, Wasmuth JD, Gilleard JS. Exploring the Gastrointestinal "Nemabiome": Deep Amplicon Sequencing to Quantify the Species Composition of Parasitic Nematode Communities. Plos One. 2015;10(12).

187. Poissant J, Gavriliuc S, Bellaw J, Redman EM, Avramenko RW, Robinson D, et al. A repeatable and quantitative DNA metabarcoding assay to characterize mixed strongyle infections in horses. Int J Parasitol. 2021;51(2-3):183-92.

188. Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, Hoffmann C, et al. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome Biol. 2012;13(7):R60.

189. Krogsgaard LR, Andersen LO, Johannesen TB, Engsbro AL, Stensvold CR, Nielsen HV, et al. Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome. Clin Transl Gastroenterol. 2018;9(6):161.

190. Gogarten JF, Calvignac-Spencer S, Nunn CL, Ulrich M, Saiepour N, Nielsen HV, et al. Metabarcoding of eukaryotic parasite communities describes diverse parasite assemblages spanning the primate phylogeny. Mol Ecol Resour. 2020;20(1):204-15.

191. Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI. How quantitative is metabarcoding: A meta-analytical approach. Mol Ecol. 2019;28(2):420-30.

192. Hall RA, Noverr MC. Fungal interactions with the human host: exploring the spectrum of symbiosis. Curr Opin Microbiol. 2017;40:58-64.

193. Han B, Takvorian PM, Weiss LM. Invasion of Host Cells by Microsporidia. Front Microbiol. 2020;11:172.

194. Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, et al. The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. Stand Genomic Sci. 2010;3(3):249-53.

195. Friant S, Ziegler TE, Goldberg TL. Primate reinfection with gastrointestinal parasites: behavioural and physiological predictors of parasite acquisition. Animal Behaviour. 2016;117:105-13.

196. Kounosu A, Murase K, Yoshida A, Maruyama H, Kikuchi T. Improved 18S and 28S rDNA primer sets for NGS-based parasite detection. Sci Rep. 2019;9(1):15789.

197. Pinol J, Senar MA, Symondson WO. The choice of universal primers and the chatacteristics of the species mixtuer determine when DNA metabarcoding can be quantitative. Molecular Ecology Notes. 2018;28:407-19.

198. Paige SB, Friant S, Clech L, Malave C, Kemigabo C, Obeti R, et al. Combining Footwear with Public Health Iconography to Prevent Soil-Transmitted Helminth Infections. Am J Trop Med Hyg. 2017;96(1):205-13.

199. Marquina D, Andersson AF, Ronquist F. New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. Mol Ecol Resour. 2019;19(1):90-104.

200. Macheriotou L, Guilini K, Bezerra TN, Tytgat B, Nguyen DT, Phuong Nguyen TX, et al. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. Ecol Evol. 2019;9(3):1211-26.

201. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C. Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. Plos One. 2014;9(2).

202. Bradley IM, Pinto AJ, Guest JS. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. Applied and Environmental Microbiology. 2016;82(19):5878-91.

203. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013;41(1):e1.

204. Loakes D. Survey and summary: The applications of universal DNA base analogues. Nucleic Acids Res. 2001;29(12):2437-47.

205. van der Loos LM, Nijland R. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Mol Ecol. 2020.

206. Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution. 2018;9(1):134-47.

207. Yeh YC, McNichol J, Needham DM, Fichot EB, Berdjeb L, Fuhrman JA. Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S. Environ Microbiol. 2021.

208. Tourlousse DM, Narita K, Miura T, Ohashi A, Matsuda M, Ohyama Y, et al. Characterization and Demonstration of Mock Communities as Control Reagents for Accurate Human Microbiome Community Measurements. Microbiol Spectr. 2022;10(2):e0191521.

209. Wang C, Zhang T, Wang Y, Katz LA, Gao F, Song W. Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. Proc Biol Sci. 2017;284(1859).

210. Verhagen LM, Incani RN, Franco CR, Ugarte A, Cadenas Y, Sierra Ruiz CI, et al. High malnutrition rate in Venezuelan Yanomami compared to Warao Amerindians and Creoles: significant associations with intestinal parasites and anemia. PLoS One. 2013;8(10):e77581.

211. Ryan SJ, Brashares JS, Walsh C, Milbers K, Kilroy C, Chapman CA. A survey of gastrointestinal parasites of olive baboons (Papio anubis) in human settlement areas of Mole National Park, Ghana. J Parasitol. 2012;98(4):885-8.

212. Friant S, Ziegler TE, Goldberg TL. Changes in physiological stress and behaviour in semi-free-ranging red-capped mangabeys (Cercocebus torquatus) following antiparasitic treatment. Proceedings of the Royal Society B-Biological Sciences. 2016;283(1835).

213. Ul-Hasan S, Bowers RM, Figueroa-Montiel A, Licea-Navarro AF, Beman JM, Woyke T, et al. Community ecology across bacteria, archaea and microbial eukaryotes in the sediment and seawater of coastal Puerto Nuevo, Baja California. PloS one. 2019;14(2):e0212355.

214. Comeau AM, Li WK, Tremblay JE, Carmack EC, Lovejoy C. Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. PLoS One. 2011;6(11):e27492.

215. Centers for Disease Control GH, Division of Parasitic Diseases and Malaria. Alphabetical Index of Parasitic Diseases 2020 [Available from: https://www.cdc.gov/parasites/az/index.html.

216. Lukes J, Stensvold CR, Jirku-Pomajbikova K, Parfrey LW. Are Human Intestinal Eukaryotes Beneficial or Commensals? Plos Pathogens. 2015;11(8).

217. Taylor MA, Coop RL, Wall R. Veterinary parasitology. Chichester, West Sussex ; Ames, Iowa: John Wiley and Sons, Inc.,; 2016.

218. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: a software environment for sequence data. Nucleic Acids Res. 2004;32(4):1363-71.

219. Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res. 2011;39(21):e145.

220. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581-3.

Acknowledgements

We gratefully acknowledge and thank Tacugama Chimpanzee Sanctuary founder Bala Amarasekaran and veterinarian Dr. Rosa Garriga for their efforts in identifying and investigating ENGS, and the entire TCS staff for collection, storage, and coordination of shipping of samples, and for their dedication to the health and conservation of chimpanzees. We also gratefully acknowledge Heather N'te Inzalaco, Travis Wentz, Sabine Pellett, and Eric Johnson at UW-Madison (WI, USA) for providing bacterial growth media and staining reagents and for advice regarding bacterial culture. We also thank Blanca Pérez and Dr. Alberto Marco Valle at UD Histologia i Anatomia Patològica, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra (Barcelona, Spain) for Gram staining of selected ENGS-affected chimpanzee paraffin-embedded tissues. We are grateful to Thomas Postler at Columbia University (NY, USA) for his knowledge of classic linguistics and guidance in naming the novel bacterium. We thank Laura Bollinger (IRF-Frederick) for editing the manuscript. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services, Department of the Navy, Department of Defense, or the institutions and companies affiliated with the authors. K.A.B.L. is an employee of the U.S. Government and L.A.E. is a military service member. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. This work was supported in part through the prime contract of Laulima Government Solutions, LLC, with the U.S. National Institute of Allergy and Infectious Diseases (NIAID) under contract no. HHSN272201800013C and Battelle Memorial Institute's former

prime contract with NIAID under contract no. HHSN272200700016I. J.H.K. performed this work as a former employee of Battelle Memorial Institute and a current employee of Tunnell Government Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under contract no. HHSN272201800013C. This research was also supported by Global Emerging Infections Surveillance (GEIS) project P0071_19_NM_06 (K.A.B.L.) and WUN A1417, the University of Wisconsin-Madison Parasitology and Vector Biology Training Program (to L.A.O.) and the University of Wisconsin-Madison John D. MacArthur Professorship Chair (to T.L.G).

Figures

Figure 1.

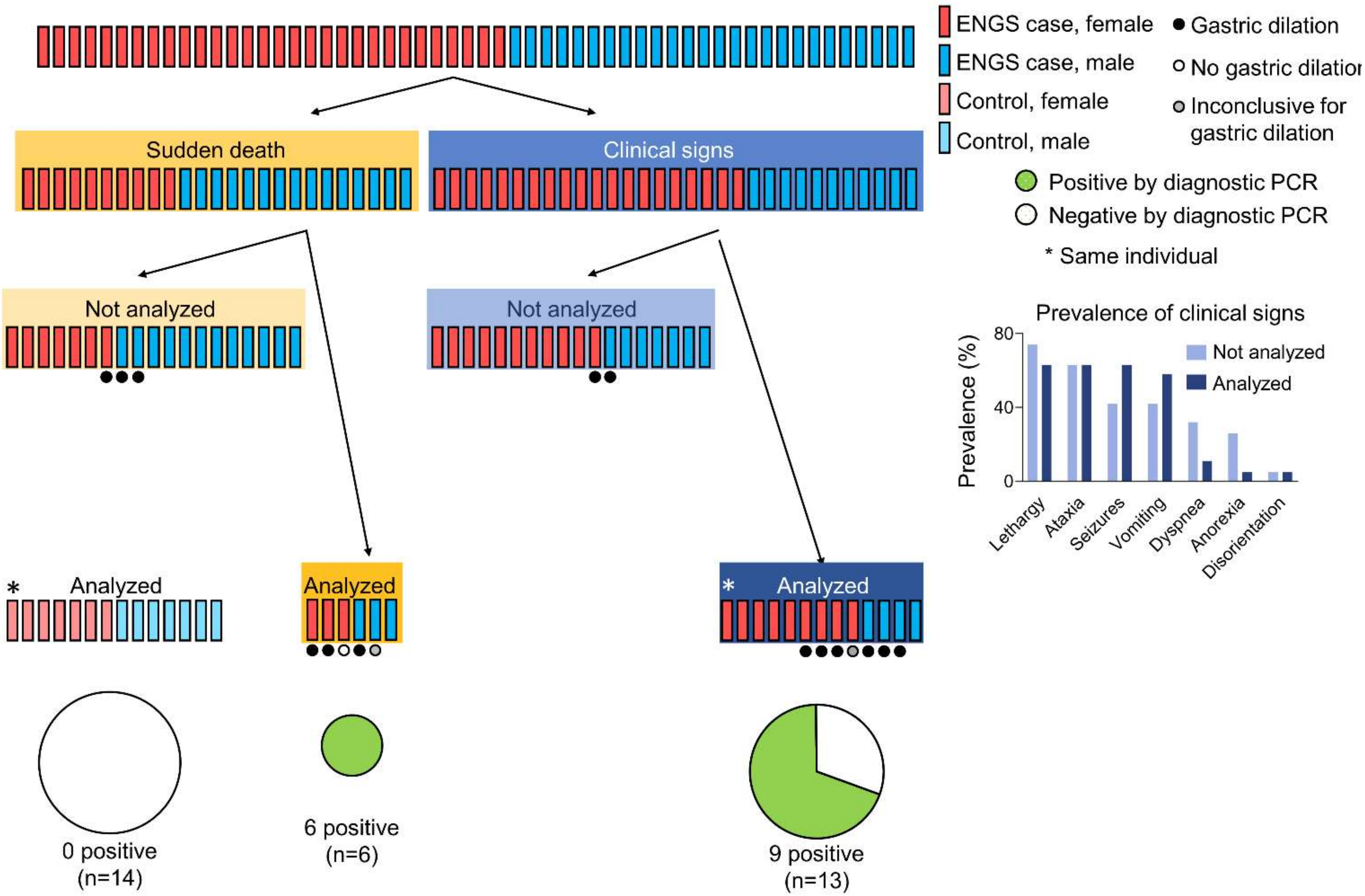


Figure 1. Graphic representation of epizootic neurologic and gastroenteric syndrome (ENGS) cases and study samples. Individual chimpanzees are represented by rectangles and colors denote sex and case/control status. Post-mortem findings regarding gastric dilation are shown for those cases where documentation was available (n=17). “Analyzed” indicates an individual from whom we obtained at least one sample used in this study (n=32) and the asterisk indicates the single individual from whom we obtained samples pre- and post-ENGS.

Figure 2.

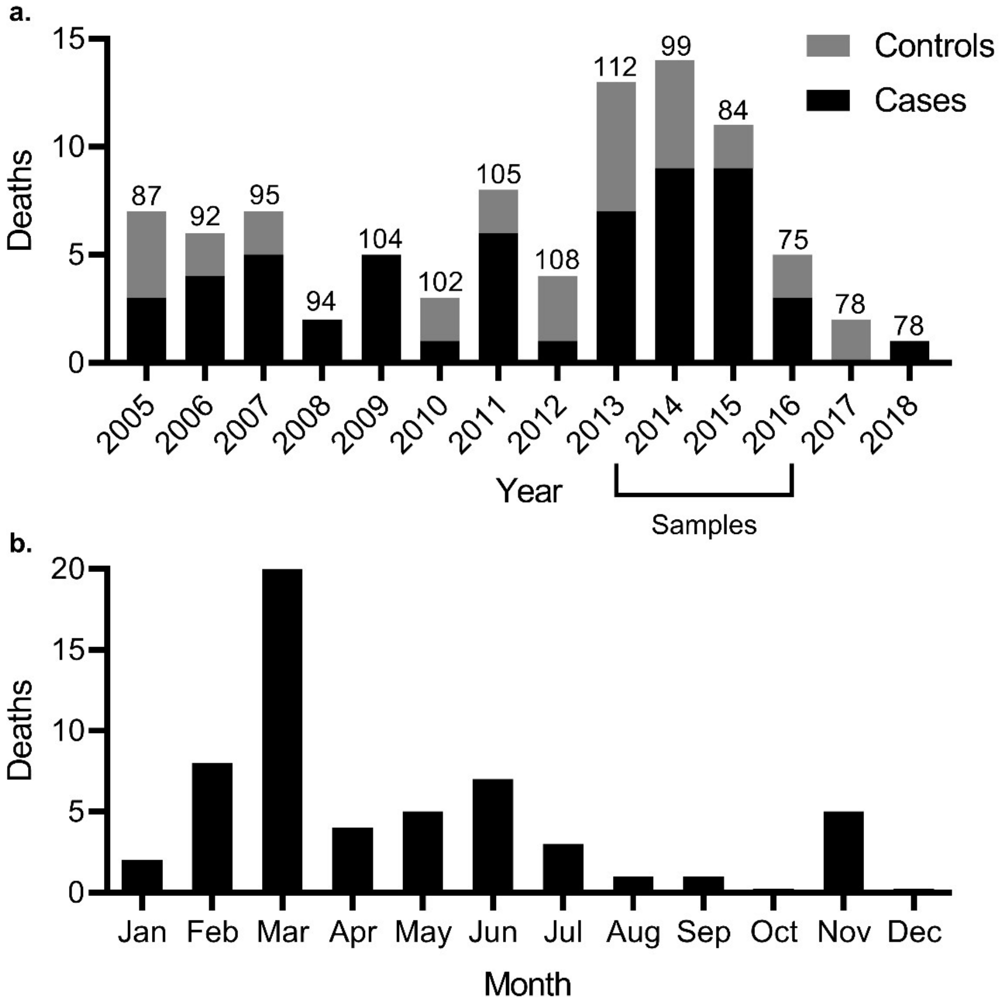


Figure 2. Deaths attributed to epizootic neurologic and gastroenteric syndrome (ENGS) at Tacugama Chimpanzee Sanctuary from 2005 through 2018. (a) Annual chimpanzee deaths from ENGS (black, overall total = 56) and other causes (gray, overall total = 32). Numbers above the bars indicate the yearly total chimpanzee population and the horizontal bracket below the x-axis denotes the period during which we obtained samples for analysis. (b) Summed totals of ENGS fatalities by month over 2005–2018 (n = 56).

Figure 3.

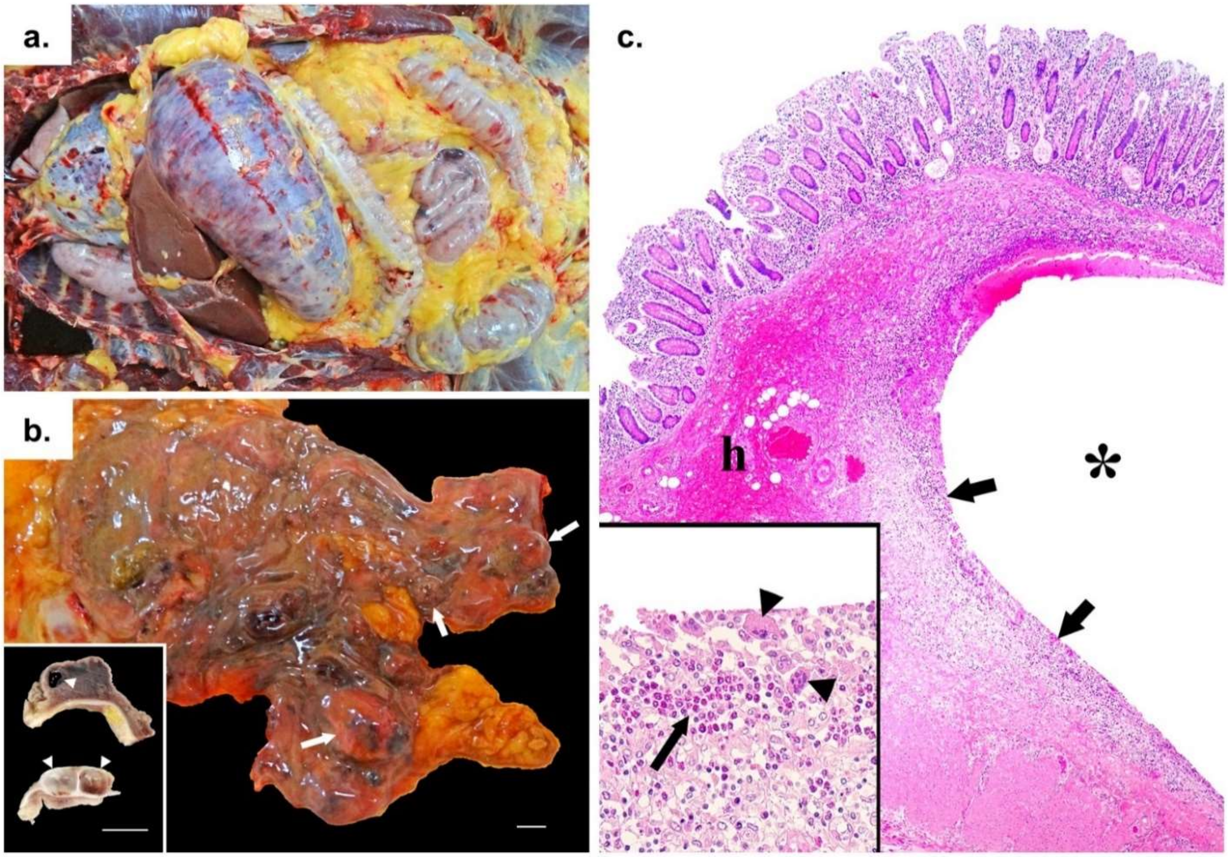


Figure 3. Gross and histopathologic images of a chimpanzee that died of epizootic neurologic and gastroenteric syndrome (ENGS). Photographic images from an adult male chimpanzee who died of ENGS showing moderate to severe gastric dilation (**a**), hemorrhagic diathesis (**a**), and emphysematous typhlocolitis (**b**; arrows point to gas-filled pockets within the caecum wall, which has reddened areas [in inset, white arrowheads point to gas bubbles in the cut surfaces of the formalin-fixed caecum]; scale bars = 1 cm). On histology, (**c**)gas-filled space (*) in the cecal submucosa, surrounded by inflammatory infiltrates (arrows) and hemorrhage (h) were visualized by hematoxylin and eosin staining (inset depicts inflammatory infiltrates, which include eosinophils [arrow] and multinucleate giant cells [arrowheads]).

Figure 4.

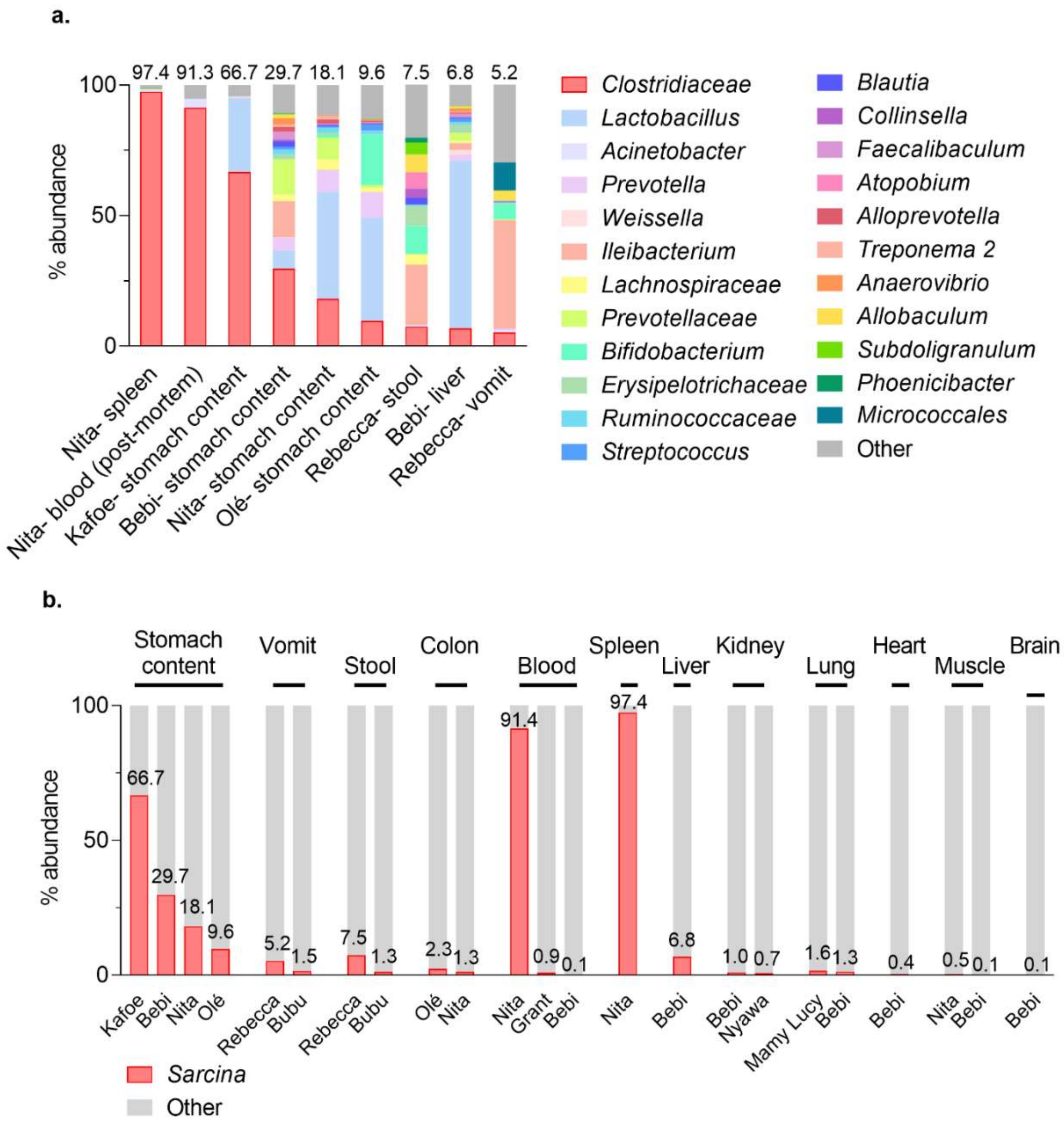


Figure 4. Bacterial 16S rDNA microbiome analysis of epizootic neurologic and gastroenteric syndrome (ENGS) case samples. (a) The percentage of total reads (n = 5,900 per sample) from 9 ENGS case samples at genus-level OTU with % reads mapped to the *Clostridiacea* OTU shown on top of each bar. (b) Percent abundance of reads from 23 ENGS case samples classified to genus-level OTUs *Sarcina* (% above each bar) and other, arranged by tissue type.

Figure 5.

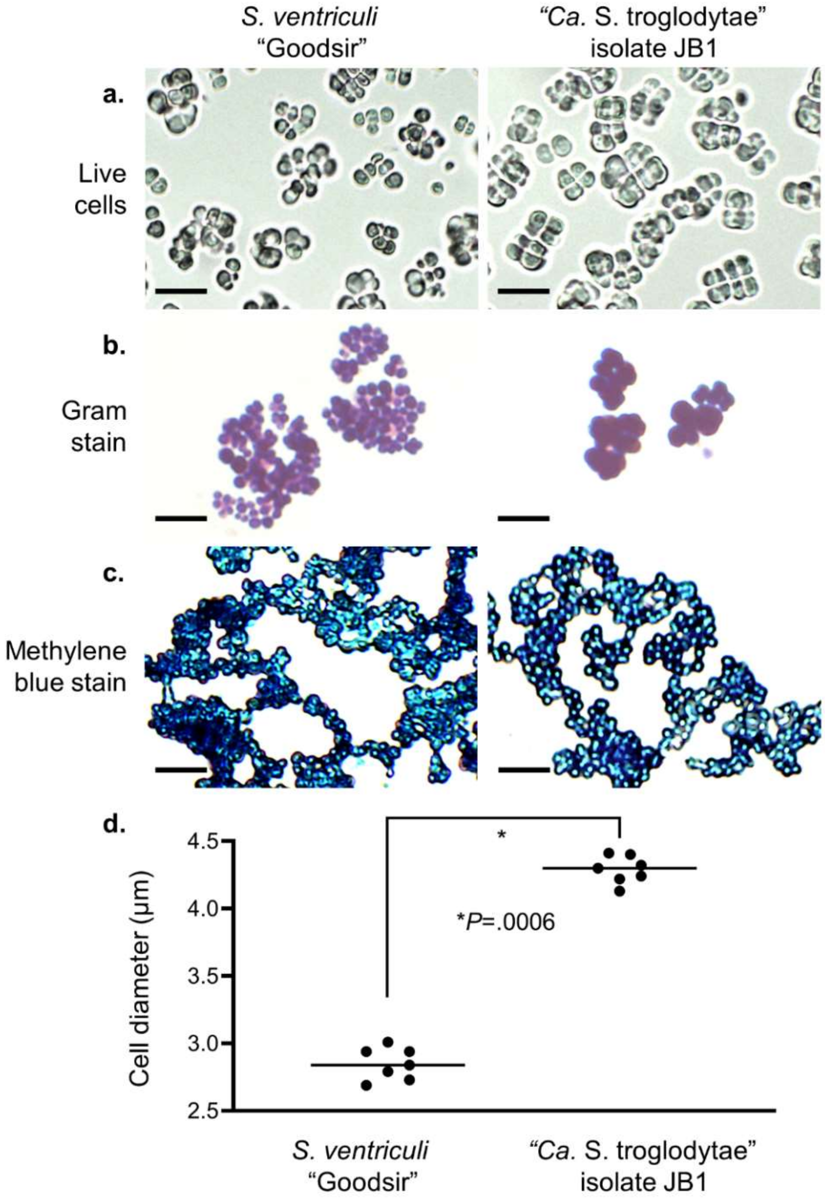


Figure 5. Comparative morphology of type strain *Sarcina ventriculi* "Goodsir" ATCC 29068 and "*Ca*. S. troglodytae" isolate JB1. (a–c) Cells were imaged live (a) or heat-fixed and stained with Gram stain (b) or methylene blue stain (c). Scale bars = 10 µm. (d) Live cell diameters were compared, with lines representing median diameters among 7 replicate bacterial colonies. *Calculated using a Mann-Whitney U test, two-tailed.

Figure 6.

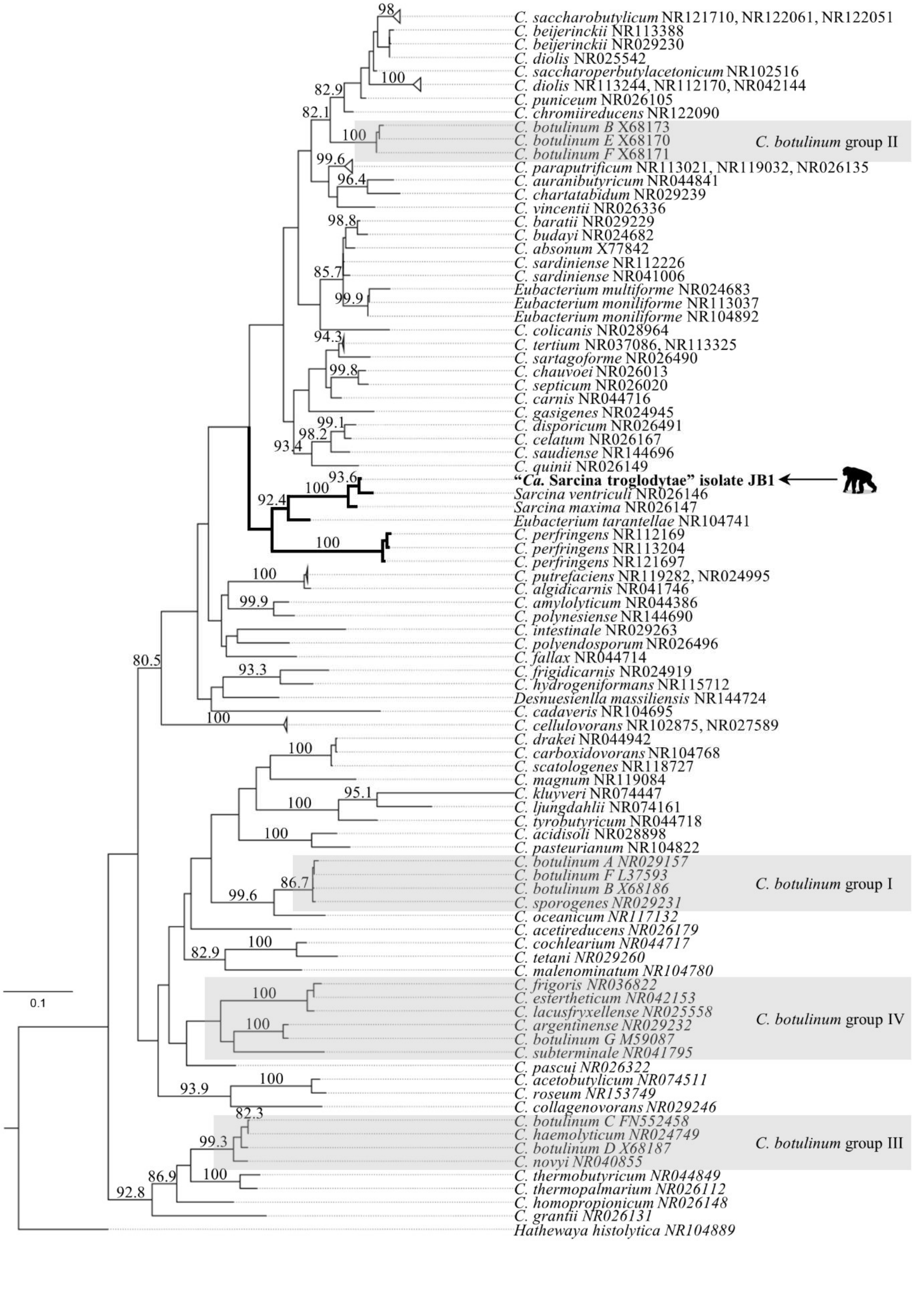

C. saccharobutylicum NR121710, NR122061, NR122051
C. beijerinckii NR113388
C. beijerinckii NR029230
C. diolis NR025542
C. saccharoperbutylacetonicum NR102516
C. diolis NR113244, NR112170, NR042144
C. puniceum NR026105
C. chromiireducens NR122090
C. botulinum B X68173
C. botulinum E X68170
C. botulinum F X68171
C. botulinum group II
C. paraputrificum NR113021, NR119032, NR026135
C. auranibutyricum NR044841
C. chartatabidum NR029239
C. vincentii NR026336
C. baratii NR029229
C. budayi NR024682
C. absonum X77842
C. sardiniense NR112226
C. sardiniense NR041006
Eubacterium multiforme NR024683
Eubacterium moniliforme NR113037
Eubacterium moniliforme NR104892
C. colicanis NR028964
C. tertium NR037086, NR113325
C. sartagoforme NR026490
C. chauvoei NR026013
C. septicum NR026020
C. carnis NR044716
C. gasigenes NR024945
C. disporicum NR026491
C. celatum NR026167
C. saudiense NR144696
C. quinii NR026149
"Ca. Sarcina troglodytae" isolate JB1
Sarcina ventriculi NR026146
Sarcina maxima NR026147
Eubacterium tarantellae NR104741
C. perfringens NR112169
C. perfringens NR113204
C. perfringens NR121697
C. putrefaciens NR119282, NR024995
C. algidicarnis NR041746
C. amylolyticum NR044386
C. polynesiense NR144690
C. intestinale NR029263
C. polyendosporum NR026496
C. fallax NR044714
C. frigidicarnis NR024919
C. hydrogeniformans NR115712
Desnuesienlla massiliensis NR144724
C. cadaveris NR104695
C. cellulovorans NR102875, NR027589
C. drakei NR044942
C. carboxidovorans NR104768
C. scatologenes NR118727
C. magnum NR119084
C. kluyveri NR074447
C. ljungdahlii NR074161
C. tyrobutyricum NR044718
C. acidisoli NR028898
C. pasteurianum NR104822
C. botulinum A NR029157
C. botulinum F L37593
C. botulinum B X68186
C. sporogenes NR029231
C. botulinum group I
C. oceanicum NR117132
C. acetireducens NR026179
C. cochlearium NR044717
C. tetani NR029260
C. malenominatum NR104780
C. frigoris NR036822
C. estertheticum NR042153
C. lacusfryxellense NR025558
C. argentinense NR029232
C. botulinum G M59087
C. subterminale NR041795
C. botulinum group IV
C. pascui NR026322
C. acetobutylicum NR074511
C. roseum NR153749
C. collagenovorans NR029246
C. botulinum C FN552458
C. haemolyticum NR024749
C. botulinum D X68187
C. novyi NR040855
C. botulinum group III
C. thermobutyricum NR044849
C. thermopalmarium NR026112
C. homopropionicum NR026148
C. grantii NR026131
Hathewaya histolytica NR104889
98
100
82.9
82.1
100
99.6
96.4
98.8
85.7
99.9
94.3
99.8
99.1
98.2
93.4
93.6
100
92.4
100
100
99.9
93.3
80.5
100
100
95.1
100
100
86.7
99.6
100
82.9
100
100
100
93.9
82.3
99.3
100
86.9
92.8
0.1

Figure 6. Maximum likelihood 16S rDNA gene phylogeny of the Clostridiaceae. The phylogeny is based on the complete 16S rDNA sequence of "*Ca*. Sarcina troglodytae" isolate JB1 (arrow and silhouette) and 98 other *Clostridia sensu stricto*, with *Hathewaya histolytica* as the outgroup. Grey boxes indicate *Clostridium botulinum* groups (62). Numbers above the branches are bootstrap values (%) based on 1,000 bootstrap replicates (only values ≥75% are shown). Scale bar indicates nucleotide substitutions per site.

Figure 7.

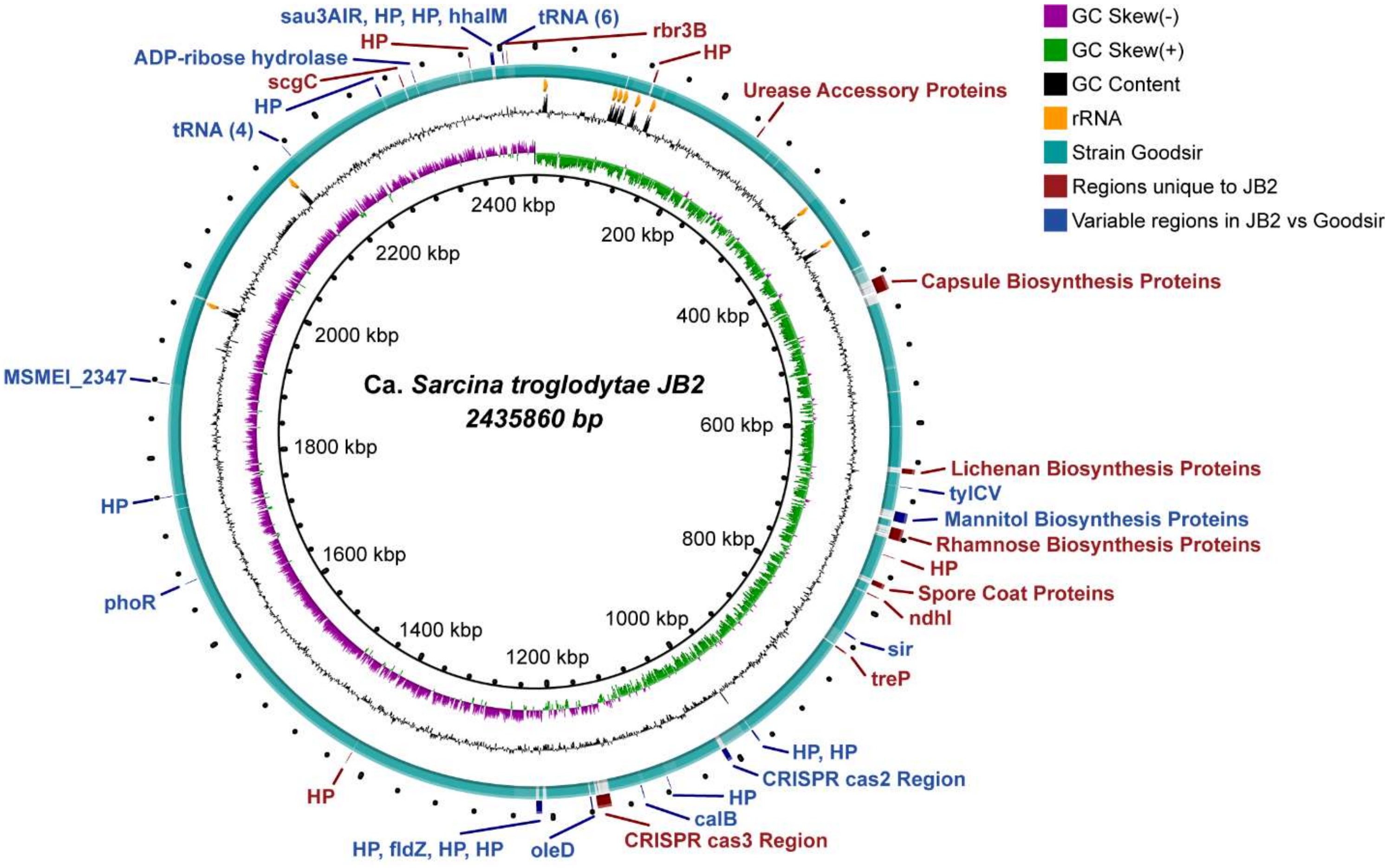
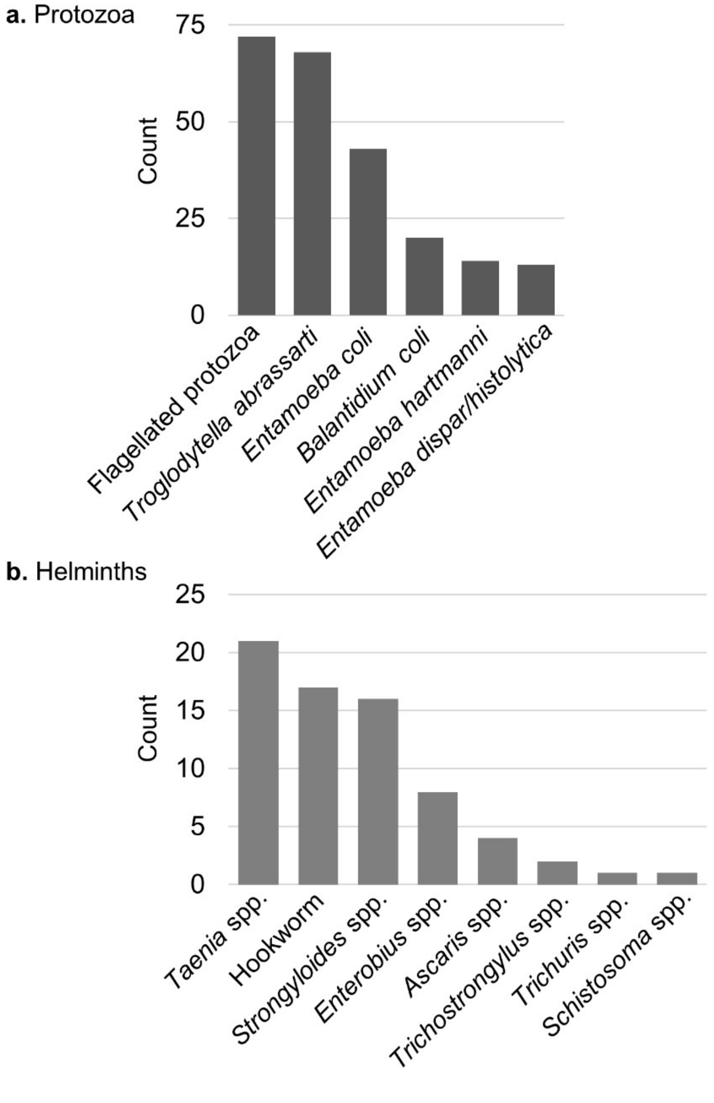


Figure 7. Whole chromosome comparison of "*Ca*. Sarcina troglodytae" isolate JB2 compared to the type strain *S. ventriculi* "Goodsir". Red: regions which are unique to JB2. Blue: different ORF sets in the same relative region in JB2 and Goodsir. Teal: 90–100% identity gradient between JB2 and Goodsir. (Goodsir chromosome sequence is based on a manually-scaffolded genome.)

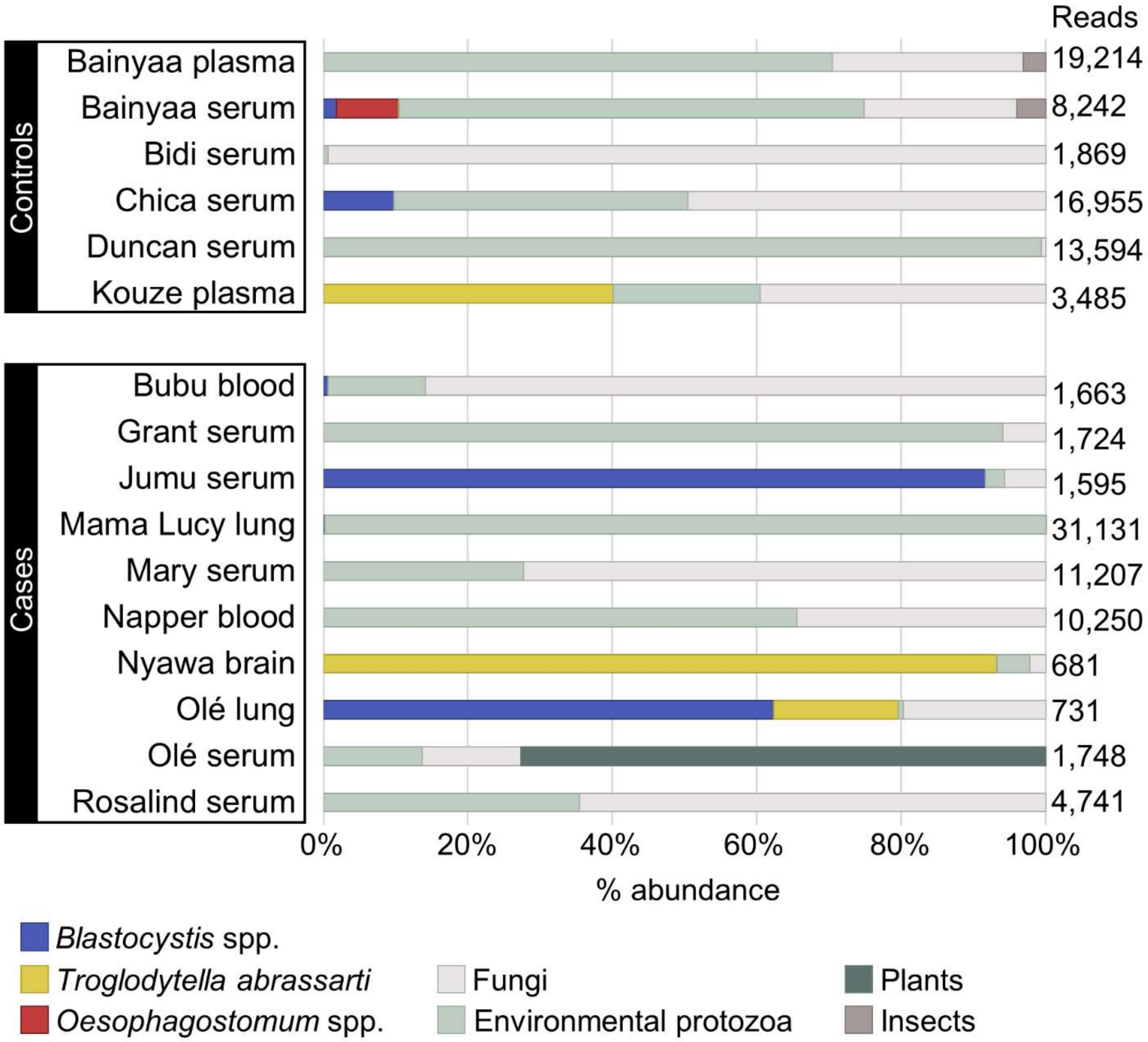
Supplementary Information

Supplementary Figure 1.



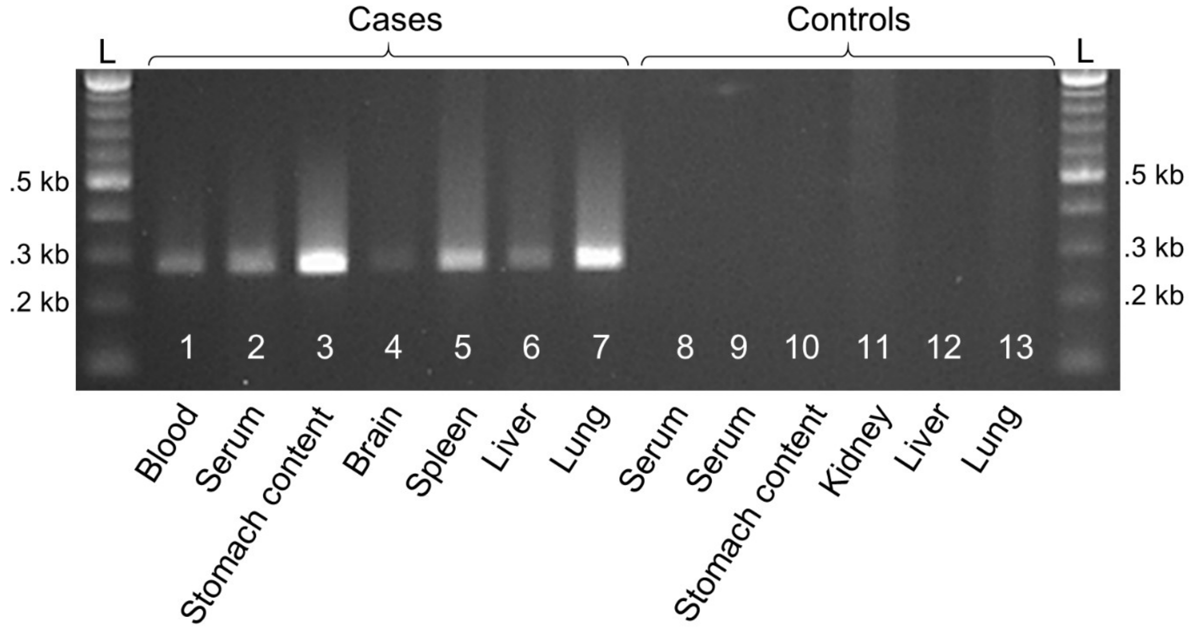
Supplementary Figure 1. Chimpanzee fecal parasitology examination results from 2005 through 2018. (a–b) Parasitology records from 17 ENGS cases and 13 controls from 2005–2018 (n = 155) are shown as a histogram in which “count” represents the number of instances a parasite was identified in an individual by any method within a 48-h period. Protozoans (a) and helminths (b) are shown separately (note different scales).

Supplementary Figure 2.



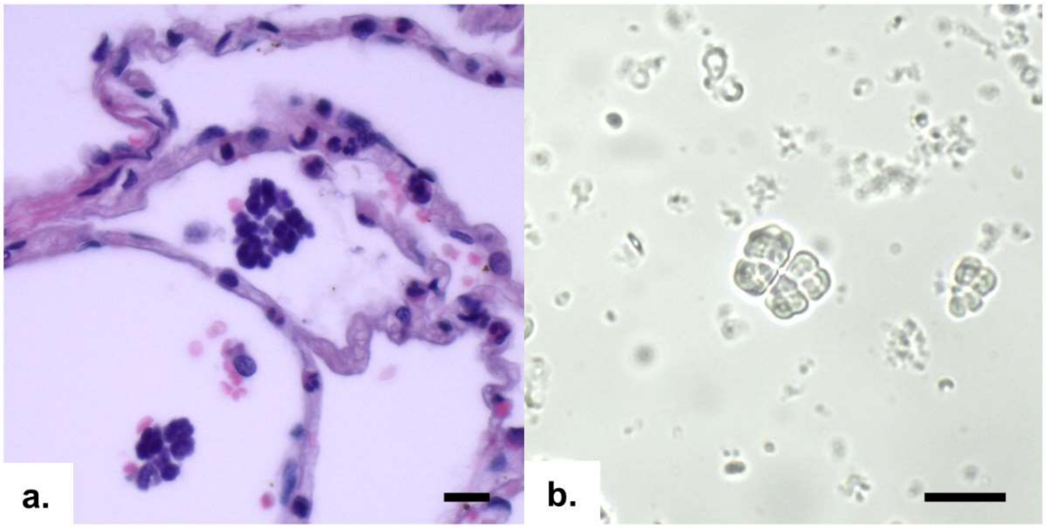
Supplementary Figure 2. Eukaryotic organisms identified with metabarcoding of the 18S rDNA gene (V9 region). Names indicate the name of the chimpanzee followed by the tissue type. The total number of reads after filtering is shown on the right, and results are expressed by OTU as percentages of total reads.

Supplementary Figure 3.



Supplementary Figure 3. Results of diagnostic PCR for "*Ca. S. troglodytae*". Case samples are "Nita's" post-mortem blood (1), "Mary's" serum (2), "Kafoe's" stomach content (3), "Joko's" brain (4), "Nita's" spleen (5), "Finda's" liver (6), and "Mama Lucy's" lung (7). Control samples are "Nita's" ante-mortem serum (8), "Mac's" serum (9), "Zeelie's" stomach content (10), "Gaura's" kidney (11), "Gaura's" liver (12), and "Gaura's" lung (13). 2-log DNA length standard shown in first and last lanes (L).

Supplementary Figure 4.



Supplementary Figure 4. Characteristic cuboidal packets of *Sarcina*-like organisms in tissues of ENGS-affected chimpanzees. (a) Basophilic packets of cells in a tetrad formation can be seen amongst and within alveoli of hematoxylin and eosin-stained lung tissue of one individual (“Jumu”). (b) Unstained brain tissue homogenate from another individual (“Joko”) contains highly refractile, cuboid packets of cells. Scale bars = 10 µm.

Supplementary Table 1. Epizootic neurologic and gastroenteric syndrome (ENGS) disease characteristics.

Number of episodes of clinical signs in ENGS cases

	Number of episodes of clinical signs				
	Low	Median	High	Mean	n
Sudden death	0	0	0	0	24
1 episode of clinical signs preceding death	1	1	1	1	21
>1 episode of clinical signs preceding death	2	2	5	2.5	11

Duration of episodes of clinical signs in ENGS cases

	Duration of episodes in days				
	Low	Median	High	Mean	n
1 episode	1	6	60	10.4	21
Overall, >1 episode	1	6	90	14.2	28
Individual average, >1 episode	4.5	8.7	22.5	14.2	11

Mortality clusters within the same enclosure

	Days between cases	
	0–2	3–4
Number of clusters	3	1

Supplementary Table 2. Samples obtained* from chimpanzees at Tacugama Chimpanzee Sanctuary from 03/2013 to 07/2016.

Name	Sex	Case or Control	Date sampled	Age	Clinical presentation	Signs (notes)	Samples
Bainyaa	F	Control	11/09/2015, 02/03/2016	8, 9	Healthy	None	Plasma, serum
Bidi	F	Control	02/06/2016	12	Healthy	None	Serum
Chica	F	Control	11/09/2015	10	Healthy	None	Serum
Duncan	M	Control	02/06/2016	5	Healthy	None	Serum
Gaura	M	Control	02/09/2016	9	Death due to trauma	None	Kidney, liver, lung, serum, stomach content
Joyce	F	Control	02/02/2016	10	Healthy	None	Serum
Kouze	M	Control	05/30/2016	13	Healthy	None	Blood, plasma
Kulay	F	Control	02/05/2016	15	Healthy	None	Serum
Linda	F	Control	02/21/2016	6	Healthy	None	Serum
Mac	M	Control	05/31/2015	8	Healthy	None	Serum
Nita	F	Control	02/03/2016	12	Healthy	None	Serum (ante-mortem)

Tom	M	Control	05/19/2015	32	Healthy	None	Plasma
Young Bruno	M	Control	06/01/2015	8	Healthy	None	Serum
Zeelie	M	Control	02/01/2014	9	Death due to trauma	None	Serum, stomach content
Bebi	F	Case	02/14/2016	12	Sudden death	Vomiting with blood	Blood, brain, cerebrospinal fluid, heart, kidney, liver, lung, muscle, serum, stomach content
Bintu	F	Case	04/29/2013	10	Signs preceding death	Ataxia, seizure, severe weakness/collapse	Blood
Bubu	M	Case	03/30/2014	12	Signs preceding death	Ataxia, seizure, severe weakness/collapse, vomiting	Blood, stool, vomit
Finda	F	Case	02/09/2015	13	Signs preceding death	Ataxia, seizure, vomiting	Kidney, liver, muscle, vomit
Grant	M	Case	06/08/2014	11	Signs preceding death	Ataxia, seizure, severe weakness/collapse	Blood, serum
Joko	M	Case	11/27/2015	13	Sudden death	Severe weakness/collapse	Blood, brain, heart, kidney, liver, lung, serum, stomach content
Julie	F	Case	05/19/2013	27	Signs preceding death	Ataxia, seizure, vomiting	Serum
Jumu	M	Case	03/14/2013	9	Signs preceding death	Seizure, severe weakness/collapse, vomiting, lethargy	Abdominal fluid, serum
Kafoe	M	Case	05/05/2016	15	Sudden death	None	Stomach content

Mama Lucy	F	Case	06/28/2014	19	Sudden death	Ataxia	Lung, spleen
Mary	F	Case	03/06/2015	12	Signs preceding death	Ataxia, seizure, vomiting, lethargy, hypokalemia, leukocytosis	Serum
Napper	F	Case	03/22/2014	14	Sudden death	Seizure, vomiting, anorexia	Blood
Nita	F	Case	07/11/2016	12	Sudden death	None	Blood (post-mortem), colon, fat, kidney, liver, muscle, pancreas, stomach content, spleen
Nyawa	F	Case	02/03/2016, 02/11/2016	13	Signs preceding death	Ataxia, seizure, severe weakness/collapse, lethargy (ran into electric fence)	Blood, brain, cerebrospinal fluid, heart, kidney, liver, lung, muscle, plasma, serum, stomach content
Olé	M	Case	05/25/2016	21	Sudden death	Severe weakness/collapse, shock, hemorrhage	Blood, colon, fat, kidney, liver, lung, muscle, pancreas, serum, spleen, stomach content, urine
Peewee	F	Case	01/15/2015	11	Signs preceding death	Ataxia, seizure, vomiting, dysphoric shouting, labored breathing	Serum
Rebecca	F	Case	04/27/2014	13	Signs preceding death	Ataxia, seizure, severe weakness/collapse, vomiting	Stool, vomit
Rosalind	F	Case	06/03/2014	14	Sudden death	Ataxia	Blood, serum

Zoyas	M	Case	03/23/2014	10	Signs preceding death	Ataxia, seizure, severe weakness/collapse, vomiting, labored breathing	Blood

*This table only includes frozen, unfixed samples that were available, amenable to shipping, and suitable for use in this study. Additional samples were collected during post-mortem examinations by veterinarians.

Supplementary Table 3. Comprehensive testing results.

Case or control	Name	Sample material	Parasite metabarcoding	Virus discovery	Bacterial metabarcoding	% *Sarcina*[c]	Diagnostic PCR	PCR-positive[d]	Culture	Isolation	Isolate sequenced
Control	Bainyaa	Plasma	Y	Y	Y		Y				
Control	Bainyaa	Serum	Y	Y	Y		Y				
Control	Bidi	Serum	Y	Y	Y		Y				
Control	Chica	Serum	Y	Y	Y		Y				
Control	Duncan	Serum	Y	Y	Y		Y				
Control	Gaura	Kidney			Y		Y				
Control	Gaura	Liver			Y		Y				
Control	Gaura	Lung			Y		Y				
Control	Gaura	Serum			Y		Y				
Control	Gaura	Stom cont			Y		Y				
Control	Joyce	Serum	Y	Y	Y		Y				
Control	Kouze	Blood			Y		Y				
Control	Kouze	Plasma	Y	Y	Y		Y				
Control	Kulay	Serum			Y		Y				
Control	Linda	Serum			Y		Y				
Control	Mac	Serum			Y		Y				
Control	Nita	Serum ant-m			Y		Y				
Control	Tom	Plasma			Y		Y				
Control	Young Bruno	Serum			Y		Y				
Control	Zeelie	Serum			Y		Y				

Control	Zeelie	Stom cont			Y		Y				
Case	Bebi	Blood			Y	0.08%[a]	Y	+	Y		
Case	Bebi	Brain			Y	0.12%	Y	+	Y		
Case	Bebi	CSF			Y		Y	+	Y		
Case	Bebi	Heart			Y	0.39%	Y	+	Y		
Case	Bebi	Kidney			Y	0.97%	Y	+	Y		
Case	Bebi	Liver			Y	6.80%	Y	+	Y	ISO	
Case	Bebi	Lung			Y	1.33%	Y	+	Y		
Case	Bebi	Muscle			Y	0.14%	Y	+			
Case	Bebi	Stom cont			Y	29.74%	Y	+	Y		
Case	Bebi	Serum			Y	Amp[b]	Y				
Case	Bintu	Blood			Y		Y	+	Y		
Case	Bubu	Blood	Y	Y	Y		Y				
Case	Bubu	Stool			Y	1.29%	Y	+	Y		
Case	Bubu	Vomit			Y	1.54%	Y	+	Y		
Case	Finda	Kidney			Y	Amp	Y	+	Y		
Case	Finda	Liver	Y	Y	Y	Amp	Y	+	Y		
Case	Finda	Muscle			Y		Y	+			
Case	Finda	Vomit			Y		Y	+			
Case	Grant	Blood			Y	0.92%	Y	+			
Case	Grant	Serum	Y	Y	Y		Y				
Case	Joko	Blood			Y	Amp	Y	+			
Case	Joko	Brain			Y	Amp	Y	+	Y	ISO	SEQ
Case	Joko	Heart			Y		Y				
Case	Joko	Kidney			Y		Y	+			
Case	Joko	Liver			Y		Y	+			
Case	Joko	Lung			Y	Amp	Y				
Case	Joko	Serum			Y		Y				

Case	Joko	Stom cont			Y		Y	+	Y
Case	Julie	Serum			Y		Y		
Case	Jumu	Abd fluid			Y		Y	+	
Case	Jumu	Serum	Y	Y	Y		Y		
Case	Kafoe	Stom cont			Y	66.70%	Y	+	Y
Case	Mama Lucy	Lung	Y	Y	Y	1.65%	Y	+	
Case	Mama Lucy	Spleen	Y	Y	Y	Amp	Y	+	
Case	Mary	Serum	Y	Y	Y		Y	+	
Case	Mary	Serum			Y		Y		
Case	Napper	Blood	Y	Y	Y		Y		
Case	Nita	Blood post-m			Y	91.38%	Y	+	Y
Case	Nita	Colon			Y	1.34%	Y	+	Y
Case	Nita	Fat			Y		Y	+	
Case	Nita	Kidney			Y	Amp	Y	+	Y
Case	Nita	Liver	Y	Y	Y		Y	+	Y
Case	Nita	Muscle			Y	0.48%	Y	+	Y
Case	Nita	Pancreas			Y		Y	+	Y
Case	Nita	Spleen	Y	Y	Y	97.38%	Y	+	Y
Case	Nita	Stom cont			Y	18.15%	Y	+	Y
Case	Nyawa	Blood			Y		Y		
Case	Nyawa	Brain	Y	Y	Y		Y	+	
Case	Nyawa	CSF	Y	Y	Y		Y		
Case	Nyawa	Heart			Y	Amp	Y		
Case	Nyawa	Kidney			Y	0.73%	Y	+	
Case	Nyawa	Liver			Y	Amp	Y	+	

Case	Nyawa	Lung			Y		Y		
Case	Nyawa	Muscle			Y		Y		
Case	Nyawa	Plasma	Y	Y	Y		Y		
Case	Nyawa	Serum			Y		Y		
Case	Nyawa	Stom cont			Y		Y	+	
Case	Olé	Blood			Y		Y		
Case	Olé	Colon			Y	2.25%	Y	+	
Case	Olé	Fat			Y		Y	+	
Case	Olé	Kidney			Y	Amp	Y	+	
Case	Olé	Liver			Y		Y	+	
Case	Olé	Lung	Y	Y	Y	Amp	Y	+	
Case	Olé	Muscle			Y		Y		
Case	Olé	Pancreas			Y		Y	+	
Case	Olé	Serum	Y	Y	Y		Y	+	
Case	Olé	Spleen			Y		Y		
Case	Olé	Stom cont			Y	9.23%	Y	+	Y
Case	Olé	Urine			Y		Y		
Case	Peewee	Serum	Y	Y	Y		Y	+	
Case	Rebecca	Stool			Y	7.46%	Y	+	Y
Case	Rebecca	Vomit			Y	4.56%	Y	+	
Case	Rosalind	Blood			Y		Y		
Case	Rosalind	Serum	Y	Y	Y		Y		
Case	Zoyas	Blood			Y		Y		

Stom cont, Stomach content; ant-m, Ante-mortem; post-m, Post-mortem; CSF, Cerebrospinal fluid; Abd fluid, Abdominal fluid

Y, Test or culture was attempted; ISO, Culture yielded at least one *Sarcina* isolate; SEQ, *Sarcina* isolate was sequenced

[a]Percent is the abundance of *Sarcina* reads after quality filtering and indicates amplification upon PCR with 16S rDNA primers followed by successful sequencing

[b] Amp indicates low quality amplification upon PCR with 16S rDNA primers and unsuccessful sequencing

[c] Blank cells indicate no amplification upon PCR with 16S rRNA primers

[d] Blank cells indicate no amplification upon PCR with diagnostic primers

Supplementary Table 4. 18S Metabarcoding statistics.

	All samples (n=24)					Low % samples (n=7)[c]		High % samples (n=17)[d]	
	No. raw reads[a]	Bases trimmed	% reads filtered for quality	% host reads post-quality filter	No. reads after filtering	% host reads post-quality filter	No. reads after filtering	% host reads post-quality filter	No. reads after filtering
Sum	1,477,507	210,385,414	21.45%[b]	83.32% [b]	129,443		161		129,282
Mean	61,563	8,766,059	25.25%	81.78%	5,393	89.86%	23.00	78.46%	7,604.82
Minimum	14	2,503	18.73%	10.33%	0	66.67%	0	10.33%	425
Median	66,144	9,384,788	22.24%	89.45%	1,694	92.59%	1	87.86%	3,485
Maximum	113,230	17,155,750	68.00%	100.00%	31,131	100.00%	150	96.02%	31,131
Standard deviation	31,862	4,642,950	10.95%	19.28%	7,975	10.62%	56.05	21.25%	8,579.33

[a] "Reads" refers to paired reads (all reads were paired)

[b] Overall % for combined data set

[c] Samples constituting < 0.5% total filtered reads

[d] Samples constituting > 0.5% total filtered reads

Supplementary Table 5. Case/Control statistical summary.

Organism ID	Organism type	Diagnostic	Prevalence in cases	Prevalence in controls	P[a]	Odds ratio	95% CI low	95% CI high
Blastocystis	Parasite	18S MB	50.00%	33.33%	0.633	2.000	0.244	16.363
Oesophagostomum	Parasite	18S MB	0.00%	16.67%	0.375	0.175	0.006	5.041
Troglodytella	Parasite	18S MB	30.00%	33.33%	>0.999	0.857	0.098	7.510
GB virus C	Virus	Virus Seq	70.59%	100.00%	0.273	0.175	0.008	3.678
Rhinovirus C	Virus	Virus Seq	0.00%	16.67%	0.261	0.105	0.004	2.959
Gemykibivirus 2	Virus	Virus Seq	41.18%	83.33%	0.155	0.140	0.013	1.474
Chimpanzee parvovirus	Virus	Virus Seq	5.88%	0.00%	>0.999	1.182	0.042	32.915
Human picobirnavirus 4	Virus	Virus Seq	23.53%	0.00%	0.309	4.333	0.202	93.159
Macaque picobirnavirus 24	Virus	Virus Seq	11.76%	16.67%	>0.999	0.667	0.049	9.022
Chimpanzee anellovirus	Virus	Virus Seq	11.76%	50.00%	0.089	0.133	0.015	1.176
Torque teno virus 4	Virus	Virus Seq	23.53%	16.67%	>0.999	1.538	0.137	17.335
Torque teno virus 23	Virus	Virus Seq	35.29%	33.33%	>0.999	1.091	0.153	7.802
Torque teno virus 14	Virus	Virus Seq	23.53%	33.33%	>0.999	0.615	0.081	4.704
Torque teno virus 16	Virus	Virus Seq	17.65%	33.33%	0.576	0.429	0.052	3.522
"*Ca*. S. troglodytae"	Bacterium	PCR	68.42%	0.00%	0.0001	56.077	2.866	1,097.182

Virus Seq, Virome shotgun sequencing; 18S MB, 18S metabarcoding

CI, Confidence interval around odds ratio

[a] Fisher's Exact test, 2-tailed

Supplementary Table 6. Viruses identified in Tacugama chimpanzees.

Virus	Accession	Genome	Closest relative (source, location, year, accession)[a]	Family[b]	Genus[b]	%ID (AA)[b]
Anellovirus	MT350347	ssDNA (circular)	Chimpanzee anellovirus (chimpanzee, Czech Republic, 2012, KT027937)	*Anelloviridae*	unclassified	90.04%
GB virus C	MT350348	ssRNA (+)	GB virus C variant troglodytes (chimpanzee, USA, 1998, AF070476)	*Flaviviridae*	*Pegivirus*	98.80%
Gemykibivirus	MT350349	ssDNA (circular)	Human associated gemykibivirus 2 (dog, Brazil, 2015, MH734235)	*Genomoviridae*	*Gemykibivirus*	99.70%
Parvovirus	MT350350	ssDNA (linear)	Parvovirus 4-like MK-2012 (chimpanzee, Cote d'Ivoire, 2002, JN798204)	*Parvoviridae*	*Protoparvovirus*	100.00%
Picobirnavirus (1)	MT350351	dsRNA (linear)	Human picobirnavirus (human, USA, 1991, AF246940)	*Picobirnaviridae*	*Picobirnavirus*	99.23%
Picobirnavirus (2)	MT350352	dsRNA (linear)	Porcine picobirnavirus (pig, India, 2013, KX374478)	*Picobirnaviridae*	unclassified	80.73%
Rhinovirus C	MT350353	ssRNA (+)	Rhinovirus C (human, USA, 2015, MG148341)	*Picornaviridae*	*Enterovirus*	98.56%
Torque teno virus (1)	MT350354	ssDNA (circular)	Torque teno virus (human, USA, 2015, KT163918)	*Anelloviridae*	*Alphatorquevirus*	59.17%
Torque teno virus (2)	MT350355	ssDNA (circular)	Torque teno virus (human, USA, 2015, KT163907)	*Anelloviridae*	unclassified	69.38%
Torque teno virus (3)	MT350356	ssDNA (circular)	Torque teno virus 14 (chimpanzee, West Africa, 2000, AB037926)	*Anelloviridae*	*Alphatorquevirus*	88.27%

Torque teno virus (4)	MT350357	ssDNA (circular)	Torque teno virus 23 (chimpanzee, Japan, 2000, NC_038342)	*Anelloviridae*	*Alphatorquevirus*	89.23%

AA, amino acid

[a] Closest match was identified by querying the viral polymerase nucleotide sequence against the NCBI’s GenBank nonredundant nucleotide database using the blastn homology searching algorithm

[b] Family, genus, and percent amino acid identity refer to the closest match to the translated viral polymerase nucleotide sequence in the NCBI’s GenBank nonredundant protein database

Supplementary Table 7. Culture conditions for isolation and propagation of "*Ca.* Sarcina troglodytae".

			Tissue preparation methods				
Medium	Source[a]	Catalog #	Minced with blades	Bead beat 5 min	80 °C 10 min	70 °C 10 min (150)	1:1 Ethanol (151)
AnaeroGRO BBE	HD	AG051	x[c]	x	x	na[d]	na
AnaeroGRO BRU	HD	AG301	x	x	x	na	na
AnaeroGRO CCFA	HD	AG501	x	x	x	na	na
AnaeroGRO EYA	HD	AG401	POS[e]	x	x	na	na
AnaeroGRO LKV	HD	AG601	x	x	x	na	na
AnaeroGRO PEA	HD	AG901	x	x	x	na	na
SVGM[f] pH 6.0, 1.5% agar	IH		POS	na	na	x	x
SVGM pH 6.0, 1.5% agar, 30 ml/l EYE	IH, HM	FD045	x	na	na	x	x
Willis and Hobbs Base, 30 ml/l EYE	HM	M1375, FD045	x	na	na	x	x
BHI Broth	SA	53286	x	na	na	x	x
Enriched CMB	BD	295982	x	x	x	na	na
AnaeroGRO PYEG	HD	AG24H	x	x	x	na	na
SVGM, pH 6.0	IH		x	na	na	na	na

SVGM, pH 2.2 (75)	IH		x	na	na	na	na
Thio w/H&K	HD	AG22H	x	x	x	na	na
TPGY	HM	M969	x	x	x	na	na

BBE, *Bacteroides* Bile Esculin Agar; BRU, *Brucella* Agar with Hemin and Vitamin K; CCFA, Cycloserine-Cefoxitin Fructose Agar; EYA, Egg Yolk Agar, Modified; LKV, Laked Blood with Kanamycin and Vancomycin Agar; PEA, Phenylethyl Alcohol Agar w/ 5% Sheep's Blood; SVGM, *Sarcina ventriculi* Growth Medium, EYE, Egg Yolk Emulsion; BHI, Brain Heart Infusion Broth; Enriched CMB, Cooked Meat Medium with Glucose, Hemin and Vitamin K; PYEG, Peptone Yeast Extract Glucose Broth; Thio w/H&K, Thioglycollate with Hemin and Vitamin K; TPGY, Tryptone Peptone Glucose Yeast Extract Broth

[a] Sources were Hardy Diagnostics, Santa Maria, CA, USA (HD); made in-house (IH); HiMedia, Mumbai, India (HM); Sigma Aldrich, St. Louis, MO, USA (SA); Becton Dickson, Franklin Lakes, NJ, USA (BD)

[b] Tissues attempted for culture were blood, brain, cerebrospinal fluid, colon, heart, kidney, liver, lung, muscle, pancreas, spleen, stomach content, stool, and vomit

[c] x: condition attempted, negative for "*Ca.* S. troglodytae"

[d] na: not assessed

[e] POS: condition yielded successful isolation, positive for "*Ca.* S. troglodytae"

[f] *Sarcina ventriculi* growth medium (ATCC medium 834) per liter: glucose 30 g, Bacto Peptone 5 g, yeast extract 5 g

Supplementary Table 8. Genetic distance between select pairs of bacterial species based on a 1,585-nucleotide alignment of the bacterial 16S rDNA gene.

	CP051754	NR026146	NR026147	NR104741	NR112169	NR113204	NR121697
	"*Ca.* S. troglodytae" JB1	*S. ventriculi*	*S. maxima*	*E. tarantellae*	*C. perfringens*	*C. perfringens*	*C. perfringens*
"*Ca.* S. troglodytae" JB1		0.687	0.932	3.409	6.529	6.331	6.242
Sarcina ventriculi	0.216		1.363	4.049	7.187	7.054	6.868
Sarcina maxima	0.262	0.319		3.584	6.844	6.882	6.81
Eubacterium tarantellae	0.497	0.517	0.529		5.632	5.461	5.348
Clostridium perfringens	0.669	0.711	0.695	0.563		0.069	0.274
Clostridium perfringens	0.651	0.695	0.695	0.548	0.071		0.136
Clostridium perfringens	0.628	0.674	0.683	0.543	0.118	0.101	

Percent pairwise nucleotide distance is shown above the diagonal (gray), with standard error of the mean shown below the diagonal

Supplementary Table 9. Sequencing read statistics after nxtrim.

Data type[a]	No. of reads	No. of base pairs	SRA accession no.
Mate Pair	941,849	182,171,554	SRR11551921
Paired End	1,513,933	393,314,190	SRR11551922

SRA, NCBI Sequence Read Archive

[a] Mate-pair and paired end reads are shown separately, all reads originated from a single sequencing run

Supplementary Table 10. Genome statistics for "*Ca*. Sarcina troglodytae" isolate JB2.

	Accession no.	Length (base pairs)	GC%	No. ORFs[a]
Chromosome	CP051754	2,435,860	27.60%	2,223
Plasmid 1	CP051755	78,882	23.40%	98
Plasmid 2	CP051756	34,634	26.80%	42
Plasmid 3	CP051757	20,671	22.70%	20
Plasmid 4	CP051758	13,674	28.50%	17
Plasmid 5	CP051759	11,514	22.60%	11
Plasmid 6	CP051760	11,304	23.90%	10
Plasmid 7	CP051761	10,963	25.00%	14
Plasmid 8	CP051762	9,993	22.50%	10
Plasmid 9	CP051763	9,792	25.00%	12
Plasmid 10	CP051764	4,566	24.30%	4

ORF, Open reading frame

[a] Identified and annotated by PATRIC

Supplementary Table 11. "*Ca*. Sarcina troglodytae" isolate JB2 genome characteristics and comparison to the type strain *S. ventriculi* "Goodsir".

	"*Ca*. S. troglodytae" JB2	*S. ventriculi* "Goodsir"
Chromosome length (base pairs)	2,435,860	2,428,884[a]
Extrachromosomal elements	10	unknown
GC%	27.60%	27.70%
# CDS	2,223	2,252[a]
# tRNA	76	91[a]
# urease ORFs	23	0[a]
ANI to "Goodsir"	98.40%	100%[a]

ANI, average nucleotide identity

[a]This analysis is based on a manually scaffolded genome

Supplementary Table 12. Literature review of human *Sarcina* cases post-1900 CE.

Ref	Year	Country	Case #	Sex	Age	*Sarcina* infection type	Symptoms and signs	Diagnosis (in addition to *Sarcina* infection)	*Sarcina* location	*Sarcina* evidence	*Sarcina* culture	Follow-up	Antibiotic treatment	Other treatment
Tolentino *et al*. (80)	2003	USA	1	M	14	GI	abdominal distension, abdominal pain	gastric perforation, gastric ulcer, hemorrhagic gastritis	abdominal fluid, gastric biopsy, omental biopsy	morphology, gas-liquid chromatography	anaerobic blood agar plate	symptomatic improvement at 5 d	gentamycin (12 d), metronidazole	
			2	M	50	GI	melena, nausea, vomiting, weight loss	esophagitis, gastric ulcer, gastritis	duodenal mass biopsy, gastric ulcer biopsy	morphology, gram stain	not attempted			
Laass *et al.* (69)	2010	Germany	3	M	3	GI	abdominal distension, anorexia, hematemesis, vomiting	Candida infection, emphysematous gastritis, gastric dilation	gastric biopsy	morphology	attempted, negative	complete recovery at 6 mo	imipenem (2 wk)	fluconazole (2 wk), omeprazole, rehydration therapy, no oral nutrition (1 wk)
Lam-Himlin *et al*. (152)	2011	USA	4	F	58	GI	abdominal pain, nausea, vomiting	gastritis, inflammatory mass in duodenum	gastric biopsy	morphology, sequencing	not attempted	gastric adenocarcinoma		partial gastrectomy for obstruction

Sauter *et al*. (153)	2013	USA	5	F	44	GI	dyspepsia	gastric hyperplastic polyps, gastric ulcer	gastric ulcer biopsy	morphology	not attempted	symptomatic improvement		metoclopramide (15 mg TID 5 mo), omeprazole (20 mg BID 5 mo), ranitidine (300 mg 5 mo)
			6	M	36	GI	abdominal pain, nausea, vomiting	none	gastric biopsy	morphology, sequencing	not attempted	no *Sarcina*		jejunostomy tube
			7	F	12	GI	dysphagia	esophagitis	gastric biopsy	morphology, sequencing	not attempted			
			8	F	46	GI	abdominal pain, abdominal spasms	duodenitis	gastric biopsy	morphology, sequencing	not attempted	continued spasms at 1 mo		
			9	M	12	GI	abdominal pain, vomiting	esophagitis, duodenitis, gastritis, H. pylori infection	duodenal biopsy, esophageal biopsy, gastric biopsy	morphology, sequencing	not attempted			
			10	F	16	GI	abdominal pain, vomiting	esophagitis, duodenitis, gastritis, H. pylori infection	duodenal biopsy, esophageal biopsy, gastric biopsy	morphology	not attempted			
Tuuminen *et al*. (105)	2013	Finland	11	F	48	blood	abdominal pain, diarrhea, fever, vomiting	septicemia	blood	morphology, gram stain, sequencing	anaerobic blood culture	asymptomatic	amoxicillin (5 d)	

Ratuapli *et al*. (101)	2013	USA	12	M	73	GI	iron-deficiency anemia	gastric polyps, gastritis	gastric biopsy	morphology	not attempted	symptomatic improvement, no *Sarcina* at 3 mo	ciprofloxacin (250 mg BID 1 wk), metronidazole (250 mg TID 1 wk)	sucralfate
Louis *et al*. (154)	2014	India	13	M	50	GI	abdominal pain, jaundice	gall stones, gastric outlet obstruction, gastritis	gastric biopsy	morphology, gram stain	not attempted	symptomatic improvement, no *Sarcina* at 3 mo	ciprofloxacin (250 mg BID 1 wk), metronidazole (250 mg TID 1 wk)	sucralfate
Kumar *et al*. (155)	2014	India	14	M	3	GI	abdominal pain, diarrhea, fever, jaundice	duodenitis, *Giardia* infection	duodenal biopsy	morphology	not attempted			
Karakus *et al*. (156)	2014	Turkey	15	M	16	GI	abdominal pain, diarrhea, nausea	*Candida* infection, celiac disease, esophagitis, gastritis	gastric biopsy	morphology	not attempted	asymptomatic		gluten-free diet

DiMaio *et al*. (157)	2014	USA	16	F	37	GI	abdominal distension, abdominal pain, anorexia, nausea	*Candida* infection, gastric hemorrhage, gastritis	gastric biopsy	morphology	not attempted	symptom resolution		omeprazole
Bhagat *et al*. (158)	2015	India	17	F	55	GI	abdominal pain, vomiting	gastric adenocarcinoma, gastric ulcer	gastric biopsy	morphology	not attempted			
Berry *et al*. (159)	2015	USA	18	F	65	GI	abdominal distension, abdominal pain, diarrhea, fatigue, melena, weakness	gastric ulcer	gastric biopsy	morphology	not attempted			proton pump inhibitor
Carrigan *et al*. (160)	2015	Canada	19	M	70s	GI	fall in hemoglobin	emphysematous esophagitis	esophageal biopsy	morphology	not attempted	*Sarcina* found at 1 yr		
Chougule *et al*. (161)	2015	India	20	M	43	lung	breathlessness, cough, fever	pulmonary gangrene	lung biopsy	morphology, gram stain	blood, sputum, and pus; all negative		vancomycin (15 mg/kg BID 4 wk), imipenem-cilastatin (1 g	amphotericin B (3 mg/kg 4 wk), posaconazole (200 mg QID)

													TID 4 wk)	
Sopha *et al*. (162)	2015	USA	21	F	32	GI	dizziness, dysphagia, headache, melena	gastric ulcer	gastric ulcer biopsy	morphology	not attempted	symptomatic improvement at 4 wk	fluoroquinolone (4 wk), metronidazole (4 wk)	omeprazole (4 wk)
Medlicott *et al.* (163)	2015	Canada	22	F	53	GI	abdominal pain, vomiting	gastritis	gastric biopsy	morphology, gram stain	not attempted	no *Sarcina* at 4 mo	metronidazole (500 mg TID)	motilium (10 mg BID)
Haroon Al Rasheed *et al*. (164)	2016	USA	23	F	57	GI	none	gastric fibrosis, gastritis	gastric ulcer biopsy	morphology, gram stain	not attempted	asymptomatic		
Bommannan *et al*. (165)	2016	India	24	M	10 mo	UT	none	urinary stricture, urinary tract infection	urine	morphology, gram stain	attempted, negative	no *Sarcina*	ciprofloxacin, metronidazole (2 wk)	urethral dilation
Canan *et al*. (71)	2017	Turkey	25	M	10	GI	abdominal distension, diarrhea, weakness	celiac disease, duodenitis	duodenal biopsy	morphology	not attempted	symptomatic improvement, no	anti-anaerobic antibiotics	gluten-free diet

												Sarcina at 7 mo	(4 wk)	
Mironova *et al*. (166)	2017	USA	26	M	85	GI	abdominal distension, abdominal pain, confusion, respiratory distress	gastric necrosis, gastric perforation, gastritis	gastric biopsy	morphology, gram stain	not attempted	tolerating a post-gastrectomy diet		total gastrectomy
de Meij *et al*. (167)	2017	Netherlands	27	F	12	GI	hematemesis, vomiting	erosive esophagitis, hemorrhagic gastritis	esophageal biopsy, gastric biopsy	morphology, IS-pro	attempted, negative	symptomatic improvement after several d, no *Sarcina* at 6 wk	ciprofloxacin (10 d), metronidazole (10 d)	omeprazole (40 mg)
			28	F	15	GI	inability to insert nasogastric tube	erosive gastritis, esophageal stenosis, gastric ulcer	gastric biopsy	morphology, IS-pro	not attempted	symptomatic improvement, no *Sarcina* at 6 wk	ciprofloxacin (10 d), metronidazole (10 d)	esophageal dilation, omeprazole (40 mg)
Behzadi *et al*.	2017	USA	29	F	65	GI	dysphagia	esophageal nodule, esophageal	esophageal biopsy,	morphology	not attempted	complicated by other	ciprofloxacin (1 wk),	esophageal dilation, proton pump inhibitor

(168)								stricture, gastric ulcer	gastric biopsy			conditions	metronidazole (1 wk)	
Rajasekar *et al.* (169)	2018	USA	30	M	65	GI	abdominal pain, confusion, cough, nausea, vomiting	emphysematous gastritis, ischemic gastritis	gastric biopsy	morphology	not attempted	hospice due to comorbidities	broad spectrum antibiotics	
Liu *et al.* (170)	2018	USA	31	F	43	GI	abdominal distension, abdominal pain, vomiting, tachycardia	gastric necrosis, ischemic gastritis, slipped gastric band	gastric biopsy	morphology	not attempted			
Elvert *et al.* (171)	2018	USA	32	F	33	blood	abdominal pain, dysuria, fever, vomiting	bacteremia, urosepsis	blood	morphology, gram stain, sequencing	anaerobic blood culture, CDC anaerobe agar plate with 5% sheep's blood	symptomatic improvement at 2 wk	levofloxacin (2 wk)	
Aggarwal *et al.* (172)	2018	India	33	F	45	GI	abdominal distension, abdominal pain, anorexia, vomiting, weight loss	*Candida* infection, gastric outlet obstruction, gastritis	gastric biopsy, gastric brush cytology	morphology	not attempted	symptomatic improvement at 1 wk	ciprofloxacin, metronidazole	proton pump inhibitor

Shetty *et al*. (173)	2018	USA	34	F	48	GI	abdominal pain, constipation, nausea, vomiting	esophagitis	esophageal brush cytology, gastric biopsy	morphology	not attempted			
Alvin *et al*. (100)	2018	USA	35	M	87	GI	abdominal pain, vomiting	emphysematous gastritis, erosive gastritis, gastric necrosis	gastric biopsy	morphology	attempted, negative	symptomatic improvement at 1 d, resolution of necrosis at 5 d	unspecified antibiotics	proton pump inhibitor
Singh *et al*. (102)	2019	USA	36	F	86	GI	abdominal distension, abdominal pain, anorexia, diarrhea, nausea, vomiting, confusion	emphysematous gastritis	gastric biopsy	morphology	not attempted	massive hematemesis, hemodynamic instability, death		
Gulati *et al*. (98)	2019	USA	37	M	54	GI	nausea, vomiting	esophagitis, gastritis	gastric biopsy	morphology	not attempted	symptomatic improvement	ciprofloxacin (250 mg BID 1 wk), metronidaz	proton pump inhibitor

													ole (250 mg TID 1 wk)	
Bortolotti *et al*. (174)	2019	France	38	M	65	blood	abdominal distension, abdominal pain, fever, hemodynamic instability	acute colonic pseudo-obstruction, septicemia	blood	morphology, gram stain, sequencing	anaerobic blood culture, anaerobic meat-yeast agar plate	resolution of fever, negative blood cultures	Piperacillin-Tazobactam (10 d)	fluconazole (10 d)
Singh *et al.* (175)	2019	Australia	39	F	14	GI	dysphagia	esophageal stricture, esophagitis, gastritis	gastric biopsy	morphology	not attempted	resolution of inflammation, no *Sarcina*	ciprofloxacin, metronidazole	
Propst *et al*. (176)	2019	USA	40	M	8	GI	vomiting	duodenal ulcer, esophagitis, gastritis	esophageal biopsy, gastric biopsy	morphology	not attempted	symptomatic improvement at 20 d, no *Sarcina* at 2 mo	ciprofloxacin (20 mg/kg BID 10 d), metronidazole (250 mg TID 10 d)	proton pump inhibitor (BID 2 wk)

			41	M	55	GI	none	esophagitis	esophageal biopsy	morphology, gram stain	not attempted	no *Sarcina* at 1.5 yr		
			42	F	65	GI	diarrhea	gastritis	duodenal biopsy, gastric biopsy	morphology	not attempted			
Dey *et al*. (177)	2019	India	43	M	52	lung	cough, fever	pulmonary TB	sputum	morphology	attempted, negative	resolution of lung lesions at 6 mo		Tuberculosis treatment
Zare *et al*. (178)	2019	USA	44	M	69	GI	abdominal distension, abdominal pain, weight loss	esophagitis, gastric outlet obstruction, gastritis	fine needle aspirate, gastric biopsy	morphology, gram stain	not attempted	symptom resolution at 3 wk, decreased inflammation/no *Sarcina* at 1,3,5 mo	ciprofloxacin (3 wk), metronidazole (3 wk)	omeprazole (3 wk)

CE, Common Era; GI, gastrointestinal; UT, urinary tract; BID, 2 times daily; TID, 3 times daily; QID, 4 times daily; TB, tuberculosis

Supplementary References

1 Crowther, J. S. *Sarcina ventriculi* in human faeces. *J Med Microbiol* **4**, 343-350, doi:10.1099/00222615-4-3-343 (1971).

2 Edwards, A. N. & McBride, S. M. Isolating and purifying *Clostridium difficile* spores. *Methods Mol Biol* **1476**, 117-128, doi:10.1007/978-1-4939-6361-4_9 (2016).

3 Vatn, S., Tranulis, M. A. & Hofshagen, M. *Sarcina* -like bacteria, *Clostridium fallax* and *Clostridium sordellii* in lambs with abomasal bloat, haemorrhage and ulcers. *J Comp Pathol* **122**, 193-200, doi:10.1053/jcpa.1999.0363 (2000).

4 Tolentino, L. E., Kallichanda, N., Javier, B., Yoshimori, R. & French, S. W. A case report of gastric perforation and peritonitis associated with opportunistic infection by *Sarcina ventriculi*. *Lab Med* **34**, 535-537, doi:Doi 10.1309/Cdff04he9fhdqpan (2003).

5 Laass, M. W. *et al.* Emphysematous gastritis caused by *Sarcina ventriculi*. *Gastrointest Endosc* **72**, 1101-1103, doi:10.1016/j.gie.2010.02.021 (2010).

6 Lam-Himlin, D. *et al. Sarcina* organisms in the gastrointestinal tract: a clinicopathologic and molecular study. *Am J Surg Pathol* **35**, 1700-1705, doi:10.1097/PAS.0b013e31822911e6 (2011).

7 Sauter, J. L. *et al.* Co-existence of *Sarcina* organisms and *Helicobacter pylori* gastritis/duodenitis in pediatric siblings. *J Clin Anat Pathol (JCAP)* **1** (2013).

8 Tuuminen, T., Suomala, P. & Vuorinen, S. *Sarcina ventriculi* in blood: the first documented report since 1872. *BMC Infect Dis* **13**, 169, doi:10.1186/1471-2334-13-169 (2013).

9 Ratuapli, S. K., Lam-Himlin, D. M. & Heigh, R. I. *Sarcina ventriculi* of the stomach: a case report. *World J Gastroenterol* **19**, 2282-2285, doi:10.3748/wjg.v19.i14.2282 (2013).

10 Louis, G. B., Singh, P. & Vaiphei, K. *Sarcina* infection. *BMJ Case Rep* **2014**, doi:10.1136/bcr-2013-201185 (2014).

11 Kumar, M., Bhagat, P., Bal, A. & Lal, S. Co-infection of *Sarcina* and *Giardia* in a child. *Oxf Med Case Reports* **2014**, 118-119, doi:10.1093/omcr/omu046 (2014).

12 Karakus, E. & Kirsaclioglu, C. T. Coincidence of celiac disease with *Sarcina* infection. *Turk J Gastroenterol* **25 Suppl 1**, 318, doi:10.5152/tjg.2014.8028 (2014).

13 DiMaio, M. A., Park, W. G. & Longacre, T. A. Gastric *Sarcina* organisms in a patient with cystic fibrosis. *Hum Path Case Rep* **1**, 45-48 (2014).

14 Bhagat, P., Gupta, N., Kumar, M., Radotra, B. D. & Sinha, S. K. A rare association of *Sarcina* with gastric adenocarcinoma diagnosed on fine-needle aspiration. *J Cytol* **32**, 50-52, doi:10.4103/0970-9371.155238 (2015).

15 Berry, A. C., Mann, S., Nakshabendi, R., Kanar, O. & Cruz, L. Gastric *Sarcina ventriculi*: incidental or pathologic? *Ann Gastroenterol* **28**, 495 (2015).

16 Carrigan, S. *et al.* Emphysematous oesophagitis associated with *Sarcina* organisms in a patient receiving anti-inflammatory therapy. *Histopath* **67**, 270-272, doi:10.1111/his.12599 (2015).

17 Chougule, A. *et al.* Pulmonary gangrene due to *Rhizopus* spp., *Staphylococcus aureus*, *Klebsiella pneumoniae* and probable *Sarcina* organisms. *Mycopath* **180**, 131-136, doi:10.1007/s11046-015-9904-3 (2015).

18 Sopha, S. C., Manejwala, A. & Boutros, C. N. *Sarcina*, a new threat in the bariatric era. *Hum Pathol* **46**, 1405-1407, doi:10.1016/j.humpath.2015.05.021 (2015).

19 Medlicott, S. A. C. *Sarcina ventricularis* complicating a patient status post vertical banded gastroplasty, a case. *J Gastroenterol Hep Res* **4**, 1481-1484 (2015).

20 Al Rasheed, M. R. & Senseng, C. G. *Sarcina ventriculi* : review of the literature. *Arch Pathol Lab Med* **140**, 1441-1445, doi:10.5858/arpa.2016-0028-RS (2016).

21 Bommannan, K., Gaspar, B. L. & Sachdeva, M. U. Pathogenic *Sarcina* in urine. *BMJ Case Rep* **2016**, doi:10.1136/bcr-2016-216991 (2016).

22 Canan, O., Ozkale, M. & Kayaseicuk, F. Duodenitis caused by *Sarcina ventriculi* in a case with Celiac disease and selective IgA deficiency. *Cukurova Med J* **42**, 766-768 (2017).

23 Mironova, M., Gobara, N., Pennell, C. P., Sherwinter, D. A. & Cimic, A. *Sarcina ventriculi:* Aa case report of gastric perforation in 85-year-old male with history of colon cancer. *J Case Rep Images Path* **3**, 20-23 (2017).

24 de Meij, T. G., van Wijk, M. P., Mookhoek, A. & Budding, A. E. Ulcerative gastritis and esophagitis in two children with *Sarcina ventriculi* infection. *Front in med* **4**, 145 (2017).

25 Behzadi, J., Modi, R. M., Goyal, K., Chen, W. & Pfeil, S. *Sarcina ventriculi* as an unknown culprit for esophageal stricturing. *ACG Case Rep J* **4**, e118, doi:10.14309/crj.2017.118 (2017).

26 Rajasekar, S., Onteddu, N. & Gupta, A. A rare case of emphysematous gastritis-*Sarcina ventriculi:* 1919. *Am J Gastroenterol* **113**, S1091 (2018).

27 Liu, L. & Gopal, P. *Sarcina ventriculi* in a patient with slipped gastric band and gastric distention. *Clin Gastroenterol Hepatol* **16**, A25-A26, doi:10.1016/j.cgh.2017.06.042 (2018).

28 Elvert, J. L., El Atrouni, W. & Schuetz, A. N. Photo Quiz: A bacterium better known by surgical pathologists than by clinical microbiologists. *J Clin Microbiol* **56** (2018).

29 Aggarwal, S. *et al.* Coinfection of *Sarcina ventriculi* and *Candida* in a patient of gastric outlet obstruction: an overloaded pyloric antrum. *Diagn Cytopathol* **46**, 876-878, doi:10.1002/dc.24048 (2018).

30 Shetty, N. U. *et al.* First documented case of *Sarcina* in esophageal brushing cytology. *Diagn Cytopathol* **46**, 886-887, doi:10.1002/dc.23986 (2018).

31 Alvin, M. & Al Jalbout, N. Emphysematous gastritis secondary to *Sarcina ventriculi*. *BMJ Case Rep* **2018**, doi:10.1136/bcr-2018-224233 (2018).

32 Singh, K. Emphysematous Gastritis Associated with *Sarcina ventriculi*. *Case Rep Gastroenterol* **13**, 207-213, doi:10.1159/000499446 (2019).

33 Gulati, R., Khalid, S., Tafoya, M. A. & McCarthy, D. Nausea and vomiting in a diabetic patient with delayed gastric emptying: do not delay diagnosis. *Dig Dis Sci* **64**, 681-684, doi:10.1007/s10620-019-05482-0 (2019).

34 Bortolotti, P. *et al. Clostridium ventriculi* bacteremia following acute colonic pseudo-obstruction: a case report. *Anaerobe* **59**, 32-34 (2019).

35 Singh, H., Weber, M. A., Low, J. & Krishnan, U. *Sarcina* in an adolescent with repaired esophageal atresia: a pathogen or a benign commensal? . *Pediatr Gastroenterol Nutr* **69**, e57, doi:10.1097/MPG.0000000000002339 (2019).

36 Propst, R. *et al. Sarcina* organisms in the upper gastrointestinal tract: A report of 3 cases with varying presentations. *Int J Surg Pathol*, 1066896919873715, doi:10.1177/1066896919873715 (2019).

37 Dey, B., Raphael, V., Banik, A. & Khonglah, Y. *Sarcina* in sputum cytology in a patient of pulmonary tuberculosis. *J Cytol* **36**, 219-221, doi:10.4103/JOC.JOC_121_18 (2019).

38 Zare, S. Y., Kubik, M. J., Savides, T. J., Hasteh, F. & Hosseini, M. A rare case of *Sarcina ventriculi* diagnosed on fine-needle aspiration. *Diagn Cytopathol* **47**, 1079-1081, doi:10.1002/dc.24270 (2019).

CHAPTER 3: **CRISPR-Cas9 mediated host signal reduction for 18S metabarcoding of eukaryotic endosymbiont assemblages**

Leah A. Owens, Mary I. Thurber, and Tony L. Goldberg

Abstract

Metabarcoding-based methods for identification of host-associated eukaryotes have the potential to revolutionize parasitology and microbial ecology, yet significant technical challenges remain. In particular, highly abundant host reads can mask the presence of less abundant target organisms, especially for sample types rich in host DNA *(*e.g. blood, tissues). Here we present a new CRISPR-Cas9 mediated approach designed to reduce host signal by selective amplicon digestion, thus enriching clinical samples for eukaryotic endosymbiont sequences during metabarcoding. Our method achieves a nearly 76 % increased efficiency in host signal reduction compared to no treatment and a nearly 60 % increased efficiency in host signal reduction compared to the most commonly used published method. Furthermore, application of our method to clinical samples allows for detection of parasite infections that would otherwise have been missed.

Keywords

CRISPR-Cas9, eukaryotic microbiome, parasitology, metabarcoding

Introduction

Metagenomic barcoding (metabarcoding) provides a high throughput alternative to traditional methods for reconstructing communities of host-associated organisms (Forsman et al., 2022). Substantial progress has been made in methods for metabarcoding bacteria and archaea (i.e., the "microbiome") (Hamady & Knight, 2009) and fungi (i.e. the "mycobiome") (Tedersoo et al., 2022), but similar progress has lagged for eukaryotic endosymbionts (defined here as all non-fungal eukaryotes residing within vertebrate hosts, spanning the continuum of parasites to commensals and including micro- and macro-organisms) (Laforest-Lapointe & Arrieta, 2018). One critical reason for this lag is that eukaryotic endosymbionts share highly similar DNA sequences with their eukaryotic hosts but usually at much lower concentration, leading to host signal interference (Lundberg et al., 2013; Sakai & Ikenaga, 2013). Polymerase chain reaction (PCR) primers designed to broadly recognize eukaryotic endosymbionts (especially metazoans, such as helminths) also often bind to and amplify host DNA (i.e., non-specific, or off-target amplification) (Belda et al., 2017; Vestheim & Jarman, 2008). Primers that recognize both host and target sequences generally detect only 10^{-3} ng parasite DNA for every ng host DNA present (Sow et al., 2019). For example, spleen tissue from mice experimentally infected via tail vein injection with *Leishmania donovoni* harbored an average of 200 promastigotes per 0.2 mg spleen tissue, resulting in an average ng parasite DNA: ng host DNA ratio of 10^{-5} (Nicolas et al., 2002; Titus et al., 1985). One "brute force" solution to this problem is ultra-deep sequencing – in other words, sequencing amplicons to great enough depth to compensate for host signal overabundance – but this approach is inefficient, costly, and biased against detecting low-abundance organisms (Alberdi et al., 2018; Belda et al., 2017). Using metabarcoding to reconstruct eukaryotic endosymbiont assemblages from feces is commonplace, but feces is so dominated by bacterial DNA that it can also interfere with detection of eukaryotes, even using primers that appear to be eukaryote-specific (Feehery et al., 2013; Jiang et al., 2020).

A reliable and efficient eukaryotic endosymbiont metabarcoding method should include a host-blocking element to enrich resulting sequences for eukaryotic endosymbiont reads in any sample type with high host DNA content (O'Rorke et al., 2012). We refer to this process as "host signal reduction" (HSR). Published HSR methods, including restriction enzyme digestion (Flaherty et al., 2018), peptide nucleic acid (PNA) clamps (Terahara et al., 2011), blocking oligonucleotides (Vestheim et al., 2011), and nested blocking primers (Mayer et al., 2020), each have advantages and disadvantages. The restriction enzyme approach, in which primers are designed such that only host amplicons contain a restriction enzyme recognition site, allowing for selective cleavage of off-target amplicons prior to sequencing (Flaherty et al., 2021), is effective, but suitable restriction sites with flanking PCR primer sites are rare or sometimes non-existent. Selective inhibition of off-target amplification during PCR is the most commonly published host signal reduction technique (Mamanova et al., 2010) and can be achieved using PNA clamps or various blocking oligonucleotides (Troedsson et al., 2008; von Wintzingerode et al., 2000). Such methods have been used in published eukaryotic endosymbiont metabarcoding studies (Hino et al., 2016; Lappan et al., 2019; Mann et al., 2020), but efficacy can be low, particularly in samples with high host biomass (Lundberg et al., 2013). Nested blocking primers were recently published for plant systems (Mayer et al., 2020) but have yet to be adapted for eukaryotic endosymbiont metabarcoding and may suffer the same drawbacks as PNA clamps and blocking oligos.

CRISPR-Cas9 (CC9) mediated removal of highly abundant off-target nucleic acids is regularly used in other sequencing-based approaches, such as chromatin structure studies (Wu et al., 2016), cancer screening (Gu et al., 2016), and plant microbiome profiling (Song & Xie, 2020). CC9 is a promising HSR method for eukaryotic endosymbiont metabarcoding because CRISPR-Cas9 nuclease activity is highly specific (Wu et al., 2014), reagents are readily available and relatively inexpensive, and the reaction components are modular such that different hosts or read types (e.g., dietary or environmental sequences) can be eliminated depending on experimental requirements. To our knowledge, however, CC9 has not been applied to HSR in the context of eukaryotic endosymbiont metabarcoding.

Here we assess the most commonly published HSR protocol for eukaryotic endosymbiont metabarcoding, the use of a PNA blocker, and demonstrate the need for a more effective approach. We design such a method based on a recombinant *Streptococcus pyogenes* CC9 system, in which vertebrate sequences are selectively targeted for cleavage and removal by host-specific short guide RNAs (sgRNAs) while leaving amplicons of interest intact for sequencing and analysis. Using *in silico* analyses, *in vitro* digests, and samples from experimentally infected animals, we show that our method is more effective than published HSR methods across various sample types. Finally, we compare the efficacy of eukaryotic endosymbiont metabarcoding for detection of known parasite infections and show that CC9 host signal reduction is necessary to detect hemoparasites in blood samples from naturally infected hosts.

Materials and Methods

Sample collection and characterization

We used archived blood, tissue, and fecal samples from wild nonhuman primates, including western chimpanzees (*Pan troglodytes verus*) from Sierra Leone and red colobus (*Procolobus rufomitratus*) from Uganda that had been collected as part of previous studies (Owens et al., 2021; Thurber et al., 2013). Appropriate permits and approvals were obtained by each research team prior to collection and shipping of samples. Blood and tissue samples from chimpanzees had been assessed for pathogenic organisms as described in Owens, et al (Owens et al., 2021). Blood samples from red colobus had been assessed for *Hepatocystis* parasites as described in Thurber, et al (Thurber et al., 2013). Blood from domestic dogs (*Canis lupis familiaris*) experimentally infected with *Dirofilaria immitis* strain "Missouri" was obtained via BEI resources (Catalog # NR-48907; Manassas, VA, USA), 20 µl was added to a glass slide and combined with two drops of 2 % formalin, and microfilaria were enumerated using phase optics at x 10 magnification. Samples were examined in triplicate and load was expressed as number of microfilariae per 20 µl of blood averaged across the three replicates. Genomic DNA (gDNA) from single hosts and parasites were obtained from in-house sample archives retained from prior studies (Owens et al, 2023).

DNA extraction and 18S V4 metabarcoding

Fecal samples were thawed on ice and homogenized by vortexing prior to transferring 0.2 g of homogenate to bead beating tubes for DNA extraction using the DNeasy PowerLyzer PowerSoil Kit (Qiagen, Hilden, Germany), with gDNA eluted in C6 buffer and stored at -20 °C. Whole blood, serum, and plasma were thawed on ice and solid tissue samples were subsampled with a sterile 3mm biopsy punch (Integra Life Sciences, Princeton, NJ, USA) while still frozen. gDNA was extracted from blood products and tissue samples using the Qiagen DNeasy Blood and Tissue kit following manufacturer's instructions, eluted in buffer AE, and stored at -20 °C.

Primers used to amplify the hypervariable 4 region (V4) of the 18S small subunit (SSU) ribosomal RNA (rRNA) gene (18S V4 hereafter) were based on published pan-eukaryotic sequences E572F and E1009R (Comeau et al., 2011), which were modified to replace individual barcodes with overhang adapters (underlined) compatible with the Nextera library preparation system (Illumina, San Diego, CA, USA): F 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCYGCGGTAATTCCAGCTC-3′ and R 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AYGGTATCTRATCRTCTTYG-3′. The most commonly published HSR method (host amplification blocking) which we used in this study was a peptide nucleic acid (PNA) mammal blocking primer (PNA Bio, Thousand Oaks, CA, USA): 5′-TCTTAATCATGGCCTCAGTT-3′ (Mann et al., 2020). Conditions for amplicon PCR with and without blocking primer were based on those described in Mann et al. (Mann et al., 2020). Resulting PCR products were cleaned using AMPure XP beads (Agencourt, Beverley, MA, USA) according to manufacturer's instructions and 5 µl was used as template in a 25-µl PCR with the Illumina Nextera XT Index Kit v2 and limited-cycle PCR with an annealing temperature of 55 °C for 10 cycles. Indexed libraries were cleaned using Agencourt AMPure XP beads and quantified using a Qubit fluorometer (ThermoFisher Scientific, Waltham, MA, USA). Libraries were sequenced on an Illumina MiSeq instrument using paired-end 300 ×300 cycle V3 chemistry.

Short guide RNA design and *in silico* screening

We used two concurrent approaches to design sgRNA sequences to target vertebrate host 18S V4: 1) the ARB 7.0 software package (Ludwig et al., 2004) with the SILVA SSU rRNA 132 Non-redundant Reference (RefNR) database (Quast et al., 2013), and 2) The Broad Institute's online CRISPick tool (https://portals.broadinstitute.org/gppx/crispick/public) (Doench et al., 2016) using human (*Homo sapiens*, NCBI RefSeq GCF_000001405.40), house mouse (*Mus musculus*, NCBI RefSeq GCF_000001635.26), domestic dog (*Canis lupus familiaris*, NCBI RefSeq GCF_000002285.5), and chimpanzee (*Pan troglodytes*, NCBI RefSeq GCF_002880755.1) genomes as input. We screened 50 candidate sgRNA sequences generated from each of these tools (n = 100 total) using SILVA TestProbe (Klindworth et al., 2013) *in silico* hybridization to the SILVA 138.1 RefNR database with maximum stringency (no mismatches between sgRNA sequence and DNA target) or allowing for a single mismatch outside of the 6-base pair "seed sequence" (**Table 1**). Resulting coverage metrics were used to choose the six sgRNA sequences that targeted the highest number of vertebrates and lowest number of eukaryotic endosymbionts for further testing: arb321, arb326, arb615 were designed in the arb software suite, and CA149, CA172, PT7.1 were designed using CRISPick. Alignments of sgRNAs with host sequences and digest maps were visualized using CLC Genomics Workbench v.20.2.4 (Qiagen, Hilden, Germany).

CRISPR-Cas9 *in vitro* digestion of representative organisms

All reagents for CC9 treatment of amplicons were components of the Alt-R CRISPR-Cas9 system (Integrated DNA Technologies, Coralville, IA, USA), based on recombinant *Streptococcus pyogenes* Cas9 nuclease, including Alt-R® S.p. Cas9 Nuclease V3, Alt-R®

CRISPR-Cas9 tracrRNA, and Alt-R® CRISPR-Cas9 crRNA. crRNA is the component containing the specific targeting sequence that, when complexed with tracrRNA, forms the functional sgRNA (see **Table 1** for sequences). Digest reactions were performed following the IDT "Alt-R CRISPR-Cas9 system – *in vitro* cleavage of target DNA with RNP complex" protocol version 2.2 using recommendations for PCR product templates of 500 – 2000 base pair lengths and 2 – 5 nM final DNA concentration per reaction.

CC9 cleavage and sgRNA specificity were initially assessed *in vitro* using a panel of genomic DNA samples extracted from single representative vertebrate hosts (n = 5) and eukaryotic endosymbionts (n = 6). Representative host organisms included: Mammal- *Ursus maritimus* (polar bear)*,* Amphibian- *Lithobates chiricahuensis* (leopard frog), Bird- *Gallus gallus* (chicken), Reptile- *Varanus varius* (monitor lizard), and Fish- *Salmo trutta* (brown trout). Representative eukaryotic endosymbiont organisms included: Protozoan- *Entamoeba histolytica* (amoeba)*,* Protozoan- *Trypanosoma brucei* (flagellate)*,* Microsporidian- *Encephalitozoon cuniculi,* Acanthocephalan- *Echinorhynchus salmonis* (spiny-headed worm)*,* Platyhelminth- *Schistosoma mansoni* (fluke)*,* Nematode- *Ascaris suum* (roundworm)*.* 18S V4 amplicon PCR was performed as described above, and resulting amplicons were used in Alt-R CRISPR-Cas9 digest reactions. Cleavage products were separated by gel electrophoresis on 1.5 % agarose gels containing .02 µg/ml ethidium bromide, visualized under ultraviolet light, and documented using a GelDoc XR imager (BioRad, Hercules, CA, USA). Successful cleavage was indicated by the presence of bands of between approximately 150 - 500 base pairs, which were discernably smaller than the full 18S V4 amplicon of approximately 700 base pairs.

Comparison of host signal reduction methods

We compared the efficacy of HSR for improving eukaryotic endosymbiont metabarcoding by performing 18S V4 library preparation in conjunction with 4 different protocols 1) CC9 digestion of amplicons using sgRNA arb321, 2) published V4 PNA mammal-blocking oligo described above [23] added to the amplicon PCR 3) both CC9 digestion and PNA mammal-blocking oligo, and 4) mock-treated control (no CRISPR-Cas9 or PNA reagents added). PCR templates consisted of gDNA extracted from chimpanzee blood, liver, lung, colon, and fecal samples (n = 3 each). 18S V4 library preparation and CC9 digests were performed as described above. For CC9 digested amplicons, uncleaved products (bands corresponding to undigested target amplicons) were excised from agarose gels using sterile razor blades and DNA was extracted from the gel matrix using a the ZymoClean Gel DNA Recovery Kit (Zymo, Irvine, CA, USA) according to manufacturer's instructions.

Optimization of CRISPR-Cas9 digest

We examined ratios of ribonucleoprotein complex (RNP) to host target DNA of 0.75:1, 1:1, and 1.25:1. CC9 treatment was also tested at two steps in the metabarcoding protocol: 1) after initial amplification and cleanup, prior to indexing PCR (requiring one digest reaction per sample) or 2) after indexing PCR, clean up and pooling of libraries (requiring one digest reaction total for the combined pool of samples). For evaluation of the effect of sgRNA targeting sequence on CC9 digest efficiency, we performed metabarcoding on chimpanzee blood samples (n = 3) using a panel of all 6 newly designed sgRNAs. We amplified 18S V4 from each sample and divided the PCR products into seven equal parts (one for each sgRNA and one for a no-treatment control) prior to library preparation followed by sequencing and quantification of host read abundance under each condition. The top three sgRNAs (arb326, CA149, PT7.1) were then tested in the same manner on a larger set of chimpanzee blood samples (n = 31).

Detection of known parasite infections in mammal blood samples

To test the effect of HSR and CC9 on detection of eukaryotic parasites in a verified infection, we performed eukaryotic endosymbiont metabarcoding on dog blood samples containing a mean of 57.8 *Dirofilaria immitis* microfilariae per 20 µl whole blood. We prepared sequencing libraries using CC9 digestion with a panel of all 6 newly designed sgRNAs, amplification with a PNA blocking oligo or mock-treated control prior to sequencing, and quantified host read abundance under each condition.

For metabarcoding of naturally infected hosts, we used whole blood samples from wild red colobus that were characterized by microscopic investigation and PCR as part of a concluded study (Thurber et al., 2013). Most samples (n = 16 of 19) had been found to contain one of two distinct lineages of the apicomplexan parasite *Hepatocystis*: species A in 12 of 16 infected hosts, and species B in 4 of 16 infected hosts (Thurber et al., 2013). We used aliquots of these same blood samples for gDNA extraction, 18S amplicon library preparation, treatment with CC9 digest or mock control, sequencing, and quantification of host read abundances.

Sequence data processing and analyses

Raw sequence reads were processed using QIIME2 v.1.9.1 (Caporaso et al., 2010). Forward and reverse reads were assembled into paired contigs using the command multiple_join_paired_ ends.py and quality filtered using the command multiple_split_libraries_fastq.py with default parameters, except for setting the Phred threshold to 30 or higher (-q 29) and minimum length to 100 bp (-l 100). Chimeras were identified with Usearch v.6.1 (Edgar et al., 2011) and removed. Reads were then assigned to OTUs using the QIIME protocol for open reference OTU picking with the command pick_open_reference_otus.py and the default UCLUST tool (v.0.2.0) (Edgar, 2010), and taxonomy was assigned to OTUs using default settings with the command assign_taxonomy.py against the SILVA database v. 132 (Quast et al., 2013).

Still-undetermined OTUs were assigned using BLAST within QIIME2 (-m blast) against the full GenBank nucleotide database (Sayers et al., 2021). OTUs constituting < 0.5 % of the total data set were removed from further analyses. Prism v.8.4.3 (GraphPad Software, Inc.) was used for plotting data and conducting statistical analyses.

Results

High host read abundance in 18S V4 metabarcoding data using a PNA clamp

18S V4 metabarcoding (Comeau et al., 2017; Mann et al., 2020) using DNA extracted from chimpanzee samples as input (n = 28) and including the mammal-blocking PNA clamp in every amplification (Mann et al., 2020) yielded a wide range of host signal relative abundances (**Figure 1a**). The percent abundance of host reads obtained was low in fecal samples (overall mean < 1 %) but high in all other sample types tested, including blood, plasma, serum, brain, liver, lung, spleen (overall mean = 93.5 %; **Supp Table 1**). Of non-fecal samples, plasma samples contained the lowest relative abundance of host reads (mean = 78.6 %) and spleen samples contained the highest (mean = 99.9 %; **Figure 1b**).

Short guide RNA design for universal eukaryotic endosymbiont enrichment

We designed six candidate vertebrate host-specific sgRNAs targeting 18S V4 (**Figure 2a**), including one fortuitously identical to the published 18S V4 mammal-blocking PNA oligo used above (arb321; **Table 1**) (Mann et al., 2020). Host DNA sequences targeted by the sgRNAs all include a protospacer adjacent motif (PAM) "NGG" required by the *Streptococcus pyogenes* Cas9 enzyme. Target sites are located centrally in 18S V4 (**Figure 2b**) such that the digestion products can be differentiated from uncleaved amplicons based on size (**Figure 2c**).

Using *in silico* hybridization to the SILVA 138 RefNR database (Quast et al., 2013) we found all six candidates to have similar mammalian complementarity (**Figure 3**), with each hybridizing to 50 % or more of mammalian sequences (mean = 66.4 %) with no mismatches and 60 % or more when allowing for a single mismatch (mean = 76.4 %). sgRNAs arb321 and arb326 were effective for mammalian hosts, but several gRNAs additionally recognized non-mammalian vertebrate groups, making them useful for a wider variety of hosts: arb615, CA149, and CA172 recognized mammal, bird, and fish sequences, while PT7.1 recognized all vertebrates (**Table 1**). All six sgRNA oligos failed to hybridize to any parasite/endosymbiont group, with the sole exception of *Trichinella pseudospiralis* (mean = 17.8 %; **Figure 3**) due to high 18S sequence similarity between *Trichinella* and mammals (mean = 45.5 % DNA identity for all sgRNA target regions combined in *Trichinella pseudospiralis* AY851258; **Supp Table 2**).

CRISPR-Cas9 *in vitro* digestion selectively cleaves target organisms

In vitro digests of 18S V4 amplicons from single representative vertebrate hosts and eukaryotic endosymbionts corresponded to SILVA TestProbe predicted coverages (**Figure 3**) and fragment sizes (**Figure 2b**). For example, CC9 digestion with the "mammal" arb321 sgRNA resulted in cleavage of mammal samples, but not amphibian, reptile, bird, or fish samples, whereas digestion with the “vertebrate” PT7.1 sgRNA resulted in cleavage of all 5 host samples including mammal, amphibian, reptile, bird, and fish (**Figure 4, left panel**). All eukaryotic endosymbiont amplicons, including protozoans (n = 2), microsporidians (n = 1), and helminths (n = 3) were unaffected by CC9 digestion using any sgRNA (**Figure 4, right panel**).

Evaluating host signal reduction methods

18S V4 metabarcoding using DNA extracted from chimpanzee samples as input (n = 15) with PNA blocker, CC9 digest, both PNA and digest, and no host signal reduction demonstrated CC9 digest to be the most effective method for enriching target read abundance for all sample types (blood, liver, lung, colon, and fecal samples; **Figure 5a**; **Supp Table 3**). Fecal samples yielded consistently low levels of host reads and were therefore not analyzed further. In tissue samples (blood, liver, lung, and colon) the overall percentage change in target (non-host) reads compared to no treatment control was significantly higher for CC9 treatment (mean 58.7 % increase in target reads, SEM 3.6 %, range 37.2 % - 79.9 %) compared to PNA (mean 1.5 %, SEM 1.3 %, range -7.1 % - 12.6 %; paired t-test: t = 6.94, df = 3, *P* = 0.0061) or combination treatment (mean -0.2 %, SEM 0.7 %, range -5.6 % - 2.9 %; paired t-test: t = 8.89, df = 3, *P* = 0.0030; **Figure 5b**).

Optimization of CRISPR-Cas9 digest

We optimized parameters of the CC9 digest by varying at the ratio of ribonucleoprotein complex to target DNA PAM sequence and found that a ratio of 1:1 was most effective at lowering host signal (**Figure 6a**). To confirm the identity of the low molecular weight (MW) bands resulting from CC9 digest of mixed samples (containing both host and parasite DNA), we compared host read abundance in the higher- and lower- MW bands to show that the cleaved products are indeed of host origin (**Figure 6b**). We also evaluated the application of the CC9 digest before and after indexing PCR. There was no significant difference in digest efficiency for CC9 treatment applied to each individual amplicon prior to library preparation compared to CC9 applied to a library pool (paired t-test: t = 0.38, df = 30, *P* = 0.18; **Figure 6c**). Because application of the digest after indexing is simpler and cheaper, we used this variation of the HSR protocol for all subsequent metabarcoding experiments.

18S V4 metabarcoding using a panel of all six newly designed sgRNAs demonstrated all sgRNAs to reduce host signal compared to mock-treated controls, with vertebrate sgRNA PT7.1 having the lowest abundance and mammal sgRNA arb321 having the highest (**Figure 6d**; **Supp Table 4**). Further testing using the three top-performing sgRNAs (arb326, CA149, and PT7.1) showed that digestion with any of the three sgRNAs significantly reduced host reads compared to no-treatment controls (arb326 compared to none, paired t-test: t = 282.2, df = 30, *P* < 0.0001; CA149 compared to none, paired t-test: t = 123.6, df = 30, *P* < 0.0001; PT7.1 compared to non, paired t-test: t = 370.3, df = 30, *P* < 0.001). There was also a small, but significant difference in signal reduction among the three sgRNAs, with CA149 being most effective (CA149 compared to arb326, paired t-test: t = 2.10, df = 30, *P* = 0.049; CA149 compared to PT7.1, paired t-test: t = 2.52, df = 30, *P* = 0.021; **Figure 6e**; **Supp Table 4**).

CRISPR-Cas9 digest validation using known parasite infections of mammals

***Dirofilaria immitis* in experimentally infected dogs** 18S V4 metabarcoding of experimentally infected dog blood samples containing *Dirofilaria immitis* microfilariae (mean 57.8 microfilariae per 20 µl whole blood) demonstrated CC9 digestion to be more effective at host signal reduction than PNA blocking oligo or mock treatment (**Figure 7a**). Specifically, CC9-digested samples yielded a higher abundance of *Dirofilaria immitis* reads (mean of 6 sgRNAs = 37.24 %, SEM =

4.38 %, range: 23.66 % - 54.59 %) than did PNA blocking oligo treatment (92.77 %) or mock control (88.96 %). Intriguingly, CC9-digested samples also recovered reads from fungi and dietary items that were not detected by the other methods (**Figure 7b**; **Supp Table 5**).

***Hepatocystis* in naturally infected red colobus** Data from wild red colobus blood samples demonstrated that, in untreated libraries, almost all reads were of host origin (mean = 99.9 %) and no hemoparasites were detected. By contrast, CC9 treated libraries from the same samples had, on average, only 42.6 % host reads, and hemoparasites were detected in 17 of 19 samples (**Figure 8**; **Supp Table 6**). These findings mirrored previous results from *Hepatocystis*-specific PCR of these same samples (Thurber et al., 2013), in which the same two species/lineages of *Hepatocystis* were detected: species A in 13 of the 17 infected samples and species B in 5 of the 17 infected samples (**Table 2**). One sample was positive by metabarcoding that was negative by PCR. Percent agreement was low between PCR and metabarcoding without HSR treatment (Cohen’s Kappa test: κ = 0.0, 95 % CI from 0.0 to 0.0) and high between PCR and metabarcoding with CC9 digest (Cohen’s Kappa test: κ = 0.855, 95 % CI from 0.581 to 1.000). Overall application of CC9 digest increased agreement with PCR 6-fold compared to no treatment (**Supp Table 7**).

Discussion

Here we show that a newly designed method using CRISPR-Cas9 and vertebrate host-targeted short guide RNAs was more effective at host signal reduction than PNA blocking or no treatment. Furthermore, in samples known from prior analyses to contain parasites, eukaryotic endosymbiont reads were rare or not detectable in samples treated with a PNA blocking primer or not treated with any HSR method. However, when the new CC9 method was applied to these same samples, the parasites were detected at high read intensities. The new CC9 method also yielded reads matching two lineages of *Hepatocystis* previously characterized in red colobus using genus specific PCR (Thurber et al., 2013).

The utility of the CC9 HSR method depends on the specificity of sgRNAs (Cho et al., 2014; Doench et al., 2016). We attempted to maximize specificity by designing sgRNAs using several complementary approaches and screening a large pool of 100 candidate oligos to identify six final sgRNA sequences. We then rigorously evaluated these six oligos *in silico* and in laboratory experiments using gDNA from individual eukaryotic organisms and from clinical samples infected with eukaryotic parasites. The consistency of our results across these conditions strongly suggests that the CC9 method is specific, effective, and robust. We note, however, that 8 % - 23 % of sequences from the nematode parasite *Trichinella pseudospiralis* were highly similar to the mammalian 18S V4 region CC9 recognition sites, reducing specificity in the case of this genus. If *Trichinella* is suspected, we recommend the use of sgRNAs CA 149 and CA 172, which have the lowest cross-reactivity. We also recommend *in silico* analysis to verify host complementarity prior to choosing a particular sgRNA.

A distinct advantage of our method is that it does not depend on the PCR primers used to amplify the 18S V4 region, as long as those primers flank the site of sgRNA complementarity. Therefore, any amplicon including the 18S V4 region is compatible with all sgRNA oligos presented here. We note that we recently published a new set of eukaryotic endosymbiont metabarcoding primers that out-performs all other published primer sets in terms of taxonomic breath, on-target amplification, and unbiased reconstruction of eukaryotic communities (Owens et al, 2023). We have examined this primer set in conjunction with the CC9 protocol described herein, and in combination the two methods achieve a similar reduction of host signal as this study (82 % less host reads compared to no treatment and 74 % compared to PNA clamp in blood samples; unpublished data). Also, because 18S V4 has the highest entropy of the hypervariable regions constituting 18S (Bradley et al., 2016; Pinol et al., 2019), and thus the highest taxonomic resolution, we expect our sgRNAs designs to stay relevant for as long as this locus remains the industry standard for eukaryotic endosymbiont metabarcoding.

Overall, we have shown that CRISPR-Cas9 digestion of amplicons reduces host signal sufficiently to allow for detection of rare eukaryotic endosymbionts and thus to increase the sensitivity and efficiency of eukaryotic endosymbiont metabarcoding. Our new method should help advance the fields of parasitology and eukaryotic community ecology, similar to how 16S prokaryote metabarcoding has facilitated the study the microbiome.

Acknowledgments

We gratefully acknowledge Lyric Bartholomay for assistance in sample acquisition, and we thank the NIH/NIAID Filariasis Research Reagent Resource Center (www.filariasiscenter.org) for providing microfilaremic dog blood samples. We thank the Uganda Wildlife Authority, the Uganda National Council for Science and Technology, and the Tacugama Chimpanzee Sanctuary in Sierra Leone for granting permission to conduct the original studies from which we obtained excess samples for this work. This research was funded by National Institutes of Health through grants 1R01AG049395-01 and 1R21AI163592-01 and by University of Wisconsin-Madison Parasitology and Vector Biology Training Grant T32AI007414.

References

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*(1), 134-147.

Belda, E., Coulibaly, B., Fofana, A., Beavogui, A. H., Traore, S. F., Gohl, D. M., . . . Riehle, M. M. (2017). Preferential suppression of *Anopheles gambiae* host sequences allows detection of the mosquito eukaryotic microbiome. *Scientific Reports*, *7*(1), 3241. https://doi.org/10.1038/s41598-017-03487-1

Bradley, I. M., Pinto, A. J., & Guest, J. S. (2016). Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. *Applied and Environmental Microbiology*, *82*(19), 5878-5891. https://doi.org/10.1128/Aem.01630-16

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335-336. https://doi.org/10.1038/nmeth.f.303

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., & Kim, J. S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res*, *24*(1), 132-141. https://doi.org/10.1101/gr.162339.113

Comeau, A. M., Douglas, G. M., & Langille, M. G. (2017). Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. *mSystems*, *2*(1). https://doi.org/10.1128/mSystems.00127-16

Comeau, A. M., Li, W. K., Tremblay, J. E., Carmack, E. C., & Lovejoy, C. (2011). Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. *PloS One*, *6*(11), e27492. https://doi.org/10.1371/journal.pone.0027492

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., . . . Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, *34*(2), 184-191. https://doi.org/10.1038/nbt.3437

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460-2461. https://doi.org/btq461 [pii]10.1093/bioinformatics/btq461

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194-2200. https://doi.org/10.1093/bioinformatics/btr381

Feehery, G. R., Yigit, E., Oyola, S. O., Langhorst, B. W., Schmidt, V. T., Stewart, F. J., . . . Quail, M. A. (2013). A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PloS One*, *8*(10), e76096.

Flaherty, B. R., Barratt, J., Lane, M., Talundzic, E., & Bradbury, R. S. (2021). Sensitive universal detection of blood parasites by selective pathogen-DNA enrichment and deep amplicon sequencing. *Microbiome*, *9*(1), 1. https://doi.org/10.1186/s40168-020-00939-1

Flaherty, B. R., Talundzic, E., Barratt, J., Kines, K. J., Olsen, C., Lane, M., . . . Bradbury, R. S. (2018). Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in blood via targeted amplicon deep sequencing. *Microbiome*, *6*(1), 164. https://doi.org/10.1186/s40168-018-0540-2

Forsman, A. M., Savage, A. E., Hoenig, B. D., & Gaither, M. R. (2022). DNA Metabarcoding Across Disciplines: Sequencing Our Way to Greater Understanding Across Scales of Biological Organization. *Integr Comp Biol*, *62*(2), 191-198. https://doi.org/10.1093/icb/icac090

Gu, W., Crawford, E. D., O'Donovan, B. D., Wilson, M. R., Chow, E. D., Retallack, H., & DeRisi, J. L. (2016). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biology*, *17*, 41. https://doi.org/10.1186/s13059-016-0904-5

Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*, *19*(7), 1141-1152. https://doi.org/10.1101/gr.085464.108

Hino, A., Maruyama, H., & Kikuchi, T. (2016). A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method. *Parasitology International*, *65*(5), 572-575.

Jiang, P., Lai, S., Wu, S., Zhao, X. M., & Chen, W. H. (2020). Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics*, *21*(1), 348. https://doi.org/10.1186/s12864-020-6749-z

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Glockner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*, *41*(1), e1. https://doi.org/10.1093/nar/gks808

Laforest-Lapointe, I., & Arrieta, M. C. (2018). Microbial Eukaryotes: a Missing Link in Gut Microbiome Studies. *mSystems*, *3*(2). https://doi.org/10.1128/mSystems.00201-17

Lappan, R., Classon, C., Kumar, S., Singh, O. P., de Almeida, R. V., Chakravarty, J., . . . Blackwell, J. M. (2019). Meta-taxonomic analysis of prokaryotic and eukaryotic gut flora in stool samples from visceral leishmaniasis cases and endemic controls in Bihar State India. *PLoS Neglected Tropical Diseases*, *13*(9), e0007444. https://doi.org/10.1371/journal.pntd.0007444

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, . . . Schleifer, K. H. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research*, *32*(4), 1363-1371. https://doi.org/10.1093/nar/gkh293

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., & Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, *10*(10), 999-1002. https://doi.org/10.1038/nmeth.2634

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., . . . Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, *7*(2), 111-118.

Mann, A. E., Mazel, F., Lemay, M. A., Morien, E., Billy, V., Kowalewski, M., . . . Wegener Parfrey, L. (2020). Biodiversity of protists and nematodes in the wild nonhuman primate gut. *Isme Journal*, *14*(2), 609-622. https://doi.org/10.1038/s41396-019-0551-4

Mayer, T., Mari, A., Almario, J., Murillo-Roos, M., Abdullah, M., Dombrowski, N., . . . Agler, M. T. (2020). Obtaining deeper insights into microbiome diversity using a simple method to block host and non-targets in amplicon sequencing. *bioRxiv*.

Nicolas, L., Prina, E., Lang, T., & Milon, G. (2002). Real-time PCR for detection and quantitation of leishmania in mouse tissues. *Journal of Clinical Microbiology*, *40*(5), 1666-1669. https://doi.org/10.1128/JCM.40.5.1666-1669.2002

O'Rorke, R., Lavery, S., & Jeffs, A. (2012). PCR enrichment techniques to identify the diet of predators. *Molecular Ecology Resources*, *12*(1), 5-17. https://doi.org/10.1111/j.1755-0998.2011.03091.x

Owens, L. A., Colitti, B., Hirji, I., Pizarro, A., Jaffe, J. E., Moittie, S., . . . Goldberg, T. L. (2021). A *Sarcina* bacterium linked to lethal disease in sanctuary chimpanzees in Sierra Leone. *Nat Commun*, *12*(1), 763. https://doi.org/10.1038/s41467-021-21012-x

Pinol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2), 407-419. https://doi.org/10.1111/mec.14776

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590-596. https://doi.org/10.1093/nar/gks1219

Sakai, M., & Ikenaga, M. (2013). Application of peptide nucleic acid (PNA)-PCR clamping technique to investigate the community structures of rhizobacteria associated with plant roots. *Journal of Microbiological Methods*, *92*(3), 281-288. https://doi.org/10.1016/j.mimet.2012.09.036

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2021). GenBank. *Nucleic Acids Res*, *49*(D1), D92-D96. https://doi.org/10.1093/nar/gkaa1023

Song, L., & Xie, K. (2020). Engineering CRISPR/Cas9 to mitigate abundant host contamination for 16S rRNA gene-based amplicon sequencing. *Microbiome*, *8*(1), 80. https://doi.org/10.1186/s40168-020-00859-0

Sow, A., Brevault, T., Benoit, L., Chapuis, M. P., Galan, M., Coeur d'acier, A., . . . Haran, J. (2019). Deciphering host-parasitoid interactions and parasitism rates of crop pests using DNA metabarcoding. *Scientific Reports*, *9*(1), 3646. https://doi.org/10.1038/s41598-019-40243-z

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., . . . Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. *Mol Ecol*, *31*(10), 2769-2795. https://doi.org/10.1111/mec.16460

Terahara, T., Chow, S., Kurogi, H., Lee, S. H., Tsukamoto, K., Mochioka, N., . . . Takeyama, H. (2011). Efficiency of peptide nucleic acid-directed PCR clamping and its application in the investigation of natural diets of the Japanese eel leptocephali. *PloS One*, *6*(11), e25715. https://doi.org/10.1371/journal.pone.0025715

Thurber, M. I., Ghai, R. R., Hyeroba, D., Weny, G., Tumukunde, A., Chapman, C. A., . . . Goldberg, T. L. (2013). Co-infection and cross-species transmission of divergent Hepatocystis lineages in a wild African primate community. *Int J Parasitol*, *43*(8), 613-619. https://doi.org/10.1016/j.ijpara.2013.03.002

Titus, R. G., Marchand, M., Boon, T., & Louis, J. A. (1985). A limiting dilution assay for quantifying Leishmania major in tissues of infected mice. *Parasite Immunology*, *7*(5), 545-555. https://doi.org/10.1111/j.1365-3024.1985.tb00098.x

Troedsson, C., Lee, R. F., Stokes, V., Walters, T. L., Simonelli, P., & Frischer, M. E. (2008). Development of a denaturing high-performance liquid chromatography method for detection of protist parasites of metazoans. *Applied and Environmental Microbiology*, *74*(14), 4336-4345.

Vestheim, H., Deagle, B. E., & Jarman, S. N. (2011). Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR. *Methods in Molecular Biology*, *687*, 265-274. https://doi.org/10.1007/978-1-60761-944-4_19

Vestheim, H., & Jarman, S. N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, *5*, 12. https://doi.org/10.1186/1742-9994-5-12

von Wintzingerode, F., Landt, O., Ehrlich, A., & Göbel, U. B. (2000). Peptide nucleic acid-mediated PCR clamping as a useful supplement in the determination of microbial diversity. *Applied and Environmental Microbiology*, *66*(2), 549-557.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., . . . Xie, W. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, *534*(7609), 652-657. https://doi.org/10.1038/nature18606

Wu, X., Kriz, A. J., & Sharp, P. A. (2014). Target specificity of the CRISPR-Cas9 system. *Quant Biol*, *2*(2), 59-70. https://doi.org/10.1007/s40484-014-0030-x

Author Contributions

L.A.O. designed the study, performed laboratory work, and collected and analyzed data. M.T. and T.L.G. obtained and analyzed samples as part of a previous study. L.A.O. and T.L.G. wrote the manuscript. All authors made substantive intellectual contributions, revised the manuscript, and approved the final draft.

Tables

Table 1. sgRNA sequences and characteristics.

ID	Target/sgRNA Seq	Orientation	PAM Seq	GC %	Seed seq	Host specificity**
arb321*	AACTGAGGCCATGATTAAGA*	sense	GGG	45	TTAAGA	Mammals
arb326	AGGCCATGATTAAGAGGGA	sense	CGG	40	GAGGGA	Mammals
arb615	GCAGCTAGGAATAATGGAAT	sense	AGG	55	TGGAAT	Mammals, Birds, Fish
PT7.1	ATTCTTGGACCGGCGCAAGA	sense	CGG	40	GCAAGA	Vertebrates
CA149	CTCAGCTAAGAGCATCGAGG	antisense	GGG	60	ATCGAGG	Mammals, Birds, Fish
CA172	TCTTAGCTGAGTGTCCCGCG	sense	GGG	55	CCCGCG	Mammals, Birds, Fish

sgRNA, short guide RNA; seq, sequence; PAM, protospacer adjacent motif; * sequence identical to V4 mammal blocking PNA oligo used in Mann et al. 2020; ** specificity to host groups determined by SILVA TestProbe *in silico* hybridization data.

Table 2. ***Hepatocystis*** **detection by PCR versus metabarcoding with and without CRISPR-Cas9 digestion.**

	PCR Positive/Negative		Metabarcoding, no treatment % reads post quality filtering		Metabarcoding, CC9 digest % reads post quality filtering	
ID #	*Hepatocystis* sp. A	*Hepatocystis* sp. B	*Hepatocystis* sp. A	*Hepatocystis* sp. B	*Hepatocystis* sp. A	*Hepatocystis* sp. B
1	Negative	Negative	0	0	0.002	0
2	Negative	Negative	0	0	0	0
3	Negative	Negative	0	0	0	0
4	Positive	Negative	0	0	0.182	0
5	Positive	Negative	0	0	0.135	0
6	Positive	Negative	0	0	0.049	0
7	Positive	Negative	0	0	0.235	0.005
8	Positive	Negative	0	0	0.215	0
9	Positive	Negative	0	0	0.164	0
10	Positive	Negative	0	0	0.083	0
11	Positive	Negative	0	0	0.302	0
12	Positive	Negative	0	0	0.123	0
13	Positive	Negative	0	0	0.278	0
14	Positive	Negative	0	0	0.36	0
15	Positive	Negative	0	0	0.047	0
16	Negative	Positive	0	0	0	0.076
17	Negative	Positive	0	0	0	0.291
18	Negative	Positive	0	0	0	0.26
19	Negative	Positive	0	0	0	0.45

Figures

Figure 1.

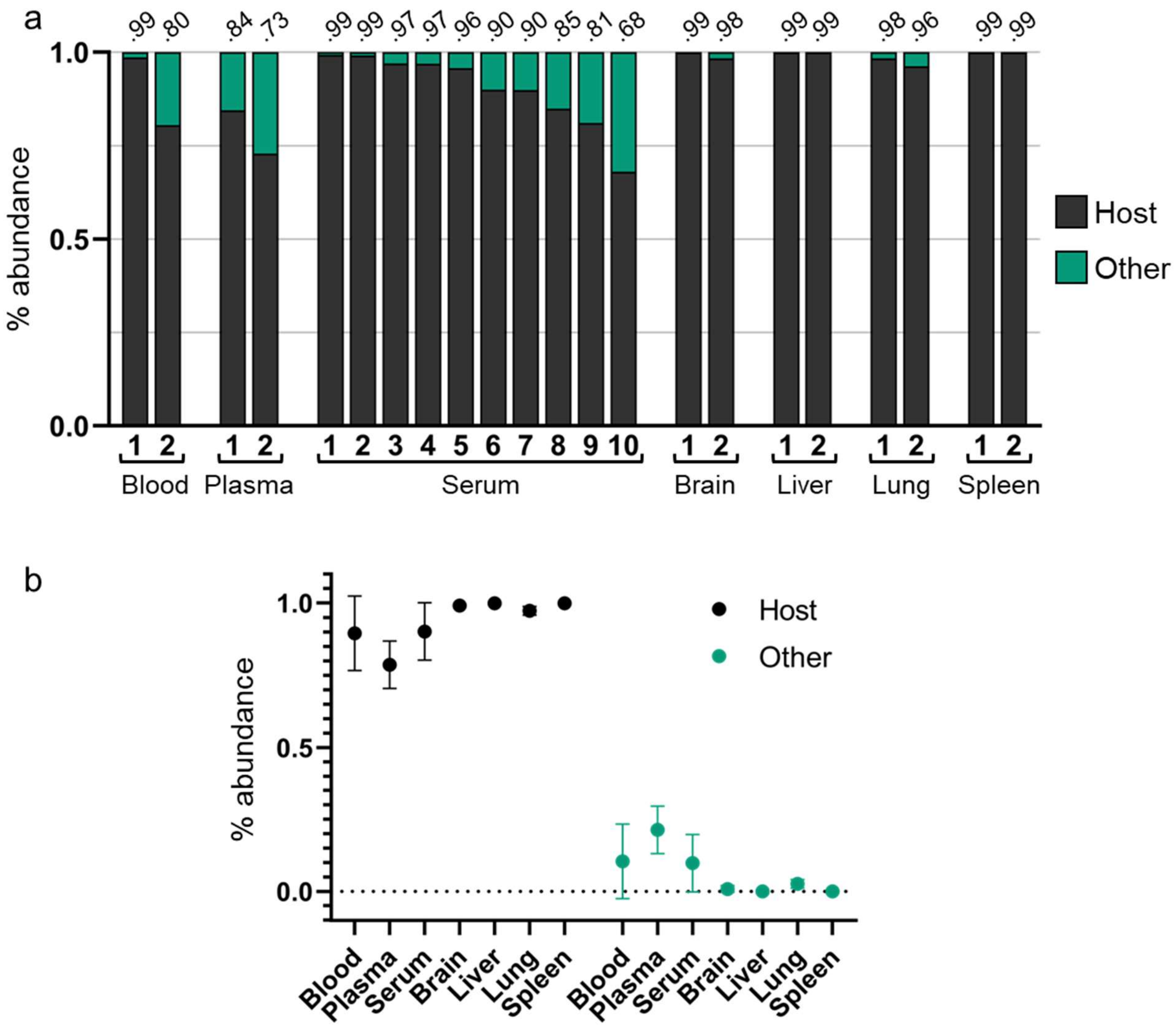


Figure 1. 18S metabarcoding with PNA mammal blocker in nonhuman primate samples. a, Percent relative abundance after quality filtering is shown for host reads (Host) and all other reads (Other). Numbers above bars represent percentage abundance of host reads. **b,** Mean relative abundance after quality filtering +/- SEM is shown for host reads (Host) and all other reads (Other). See **Supp Table 1** for source data.

Figure 2.

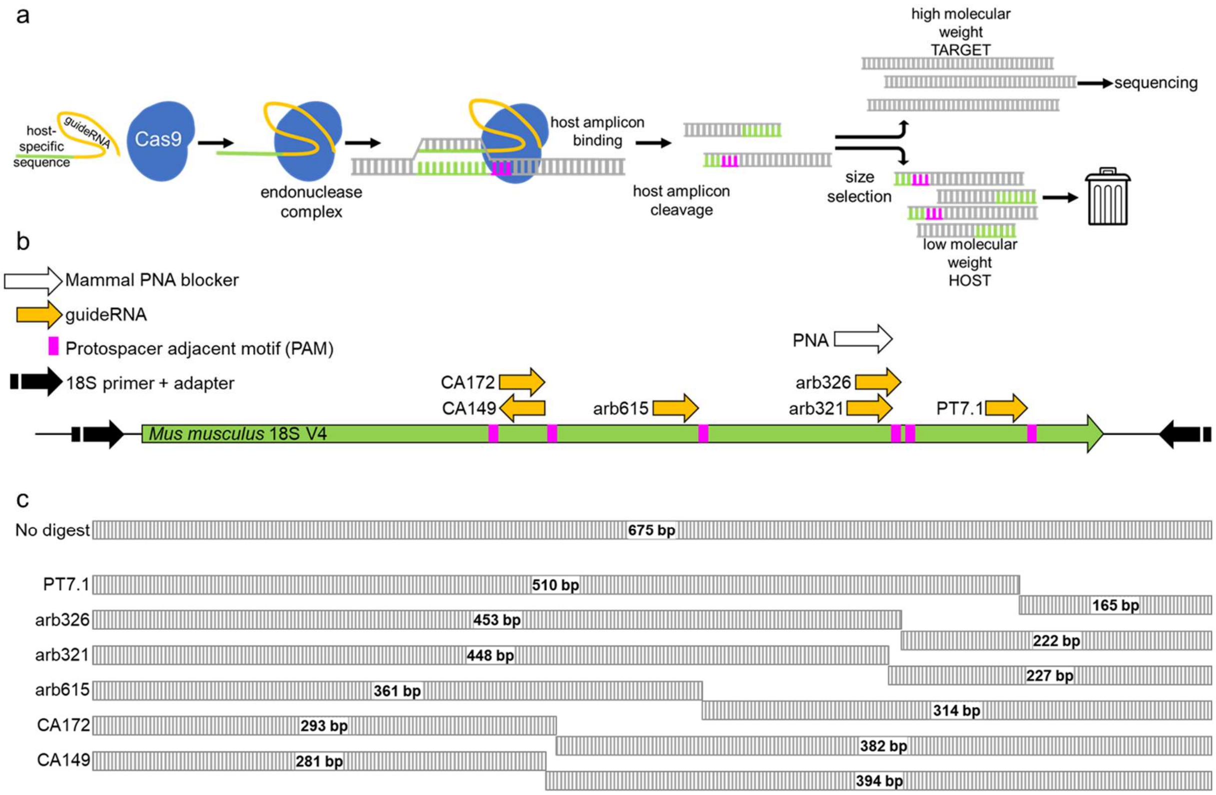
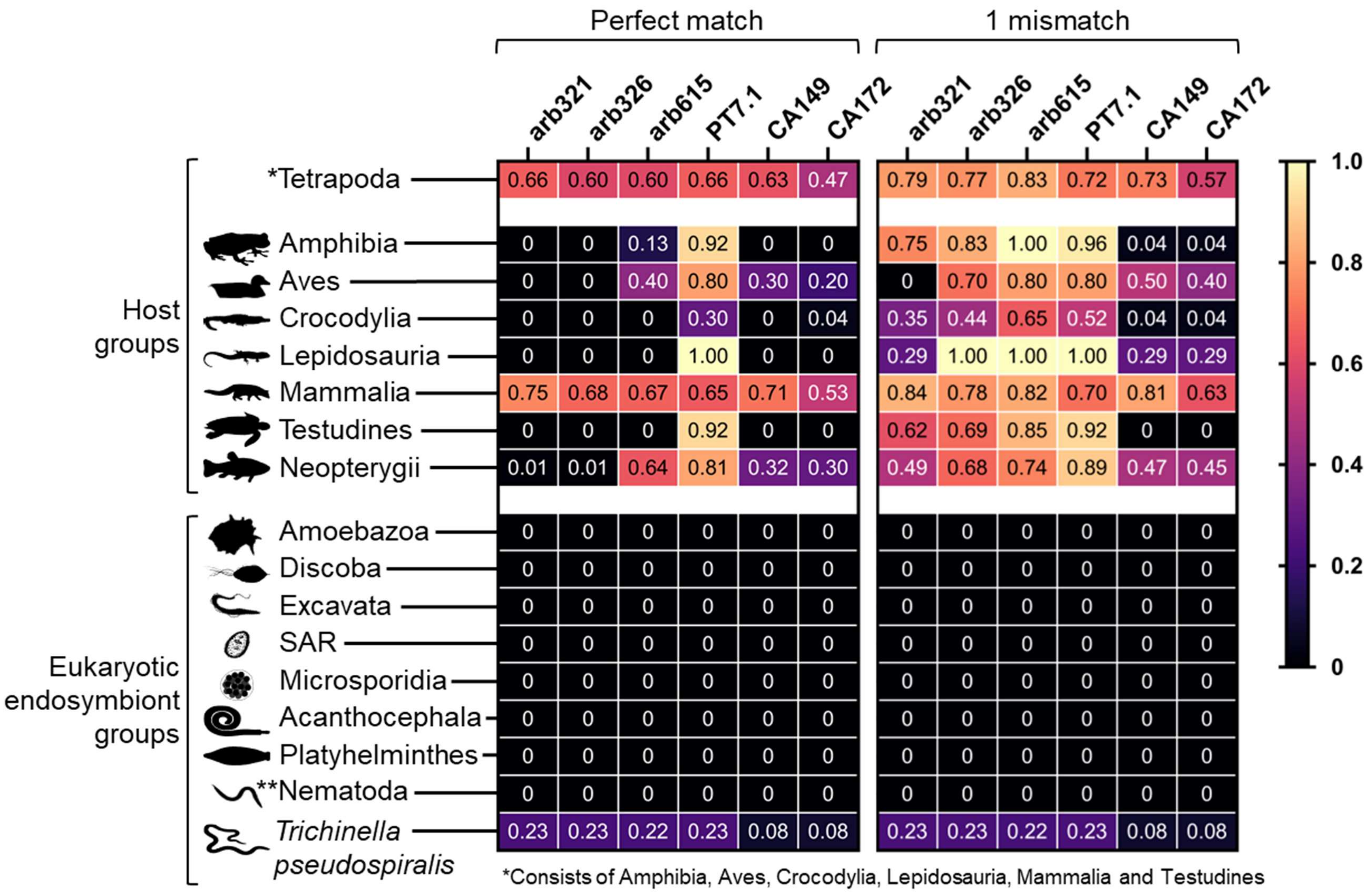


Figure 2. Overview of CRISPR-Cas9 host digestion method. a, Schematic of steps in CRISPR-Cas9 *in vitro* digestion targeting host amplicons. **b,** Map of representative mammal 18S rRNA gene (green region) from the house mouse *Mus musculus* (GenBank NR_003278) with locations of 18S amplicons primers (black arrows), newly designed short guide RNA (sgRNA) sequences (yellow arrows), and published PNA mammal blocker (white arrow). Protospacer adjacent motifs (PAMs) within the host 18S sequence are shown in pink. sgRNAs must bind next to a PAM sequence, and binding determines the location of cleavage by the Cas9 ribonucleoprotein complex. **c,** Schematic of digestion products of mouse 18S V4 amplicons using sgRNAs to target various sites. Topmost fragment (no digest) is the full-length host amplicon. Labels to the left are sgRNA names. See **Table 1** for sgRNA and PAM sequences.

Figure 3.



	Perfect match						1 mismatch					
	arb321	arb326	arb615	PT7.1	CA149	CA172	arb321	arb326	arb615	PT7.1	CA149	CA172
*Tetrapoda	0.66	0.60	0.60	0.66	0.63	0.47	0.79	0.77	0.83	0.72	0.73	0.57
Amphibia	0	0	0.13	0.92	0	0	0.75	0.83	1.00	0.96	0.04	0.04
Aves	0	0	0.40	0.80	0.30	0.20	0	0.70	0.80	0.80	0.50	0.40
Crocodylia	0	0	0	0.30	0	0.04	0.35	0.44	0.65	0.52	0.04	0.04
Lepidosauria	0	0	0	1.00	0	0	0.29	1.00	1.00	1.00	0.29	0.29
Mammalia	0.75	0.68	0.67	0.65	0.71	0.53	0.84	0.78	0.82	0.70	0.81	0.63
Testudines	0	0	0	0.92	0	0	0.62	0.69	0.85	0.92	0	0
Neopterygii	0.01	0.01	0.64	0.81	0.32	0.30	0.49	0.68	0.74	0.89	0.47	0.45
Amoebazoa	0	0	0	0	0	0	0	0	0	0	0	0
Discoba	0	0	0	0	0	0	0	0	0	0	0	0
Excavata	0	0	0	0	0	0	0	0	0	0	0	0
SAR	0	0	0	0	0	0	0	0	0	0	0	0
Microsporidia	0	0	0	0	0	0	0	0	0	0	0	0
Acanthocephala	0	0	0	0	0	0	0	0	0	0	0	0
Platyhelminthes	0	0	0	0	0	0	0	0	0	0	0	0
**Nematoda	0	0	0	0	0	0	0	0	0	0	0	0
Trichinella pseudospiralis	0.23	0.23	0.22	0.23	0.08	0.08	0.23	0.23	0.22	0.23	0.08	0.08

*Consists of Amphibia, Aves, Crocodylia, Lepidosauria, Mammalia and Testudines

**Excluding *Trichinella pseudospiralis*

Figure 3. Short guide RNA complementarity to host and eukaryotic endosymbiont groups. Percent coverage of the SILVA 138 Ref NR database is shown with numbers and color scale. **left panel**, SILVA TestProbe with the most stringent settings (no mismatches, no N's considered as matches). **right panel**, SILVA TestProbe allowing for a single mismatch outside of the conserved "seed" sequence. Taxonomic groups containing non-target "Host" groups and target "Eukaryotic endosymbiont" groups are shown with representative organism icons to the left of the heatmap. Tetrapoda includes the "Host" groups Amphibia, Aves, Crocodylia, Lepidosauria, Mammalia and Testudines. Nematoda includes all nematode accessions other than *Trichinella pseudospiralis*. See **Table 1** for sgRNA sequences.

Figure 4.

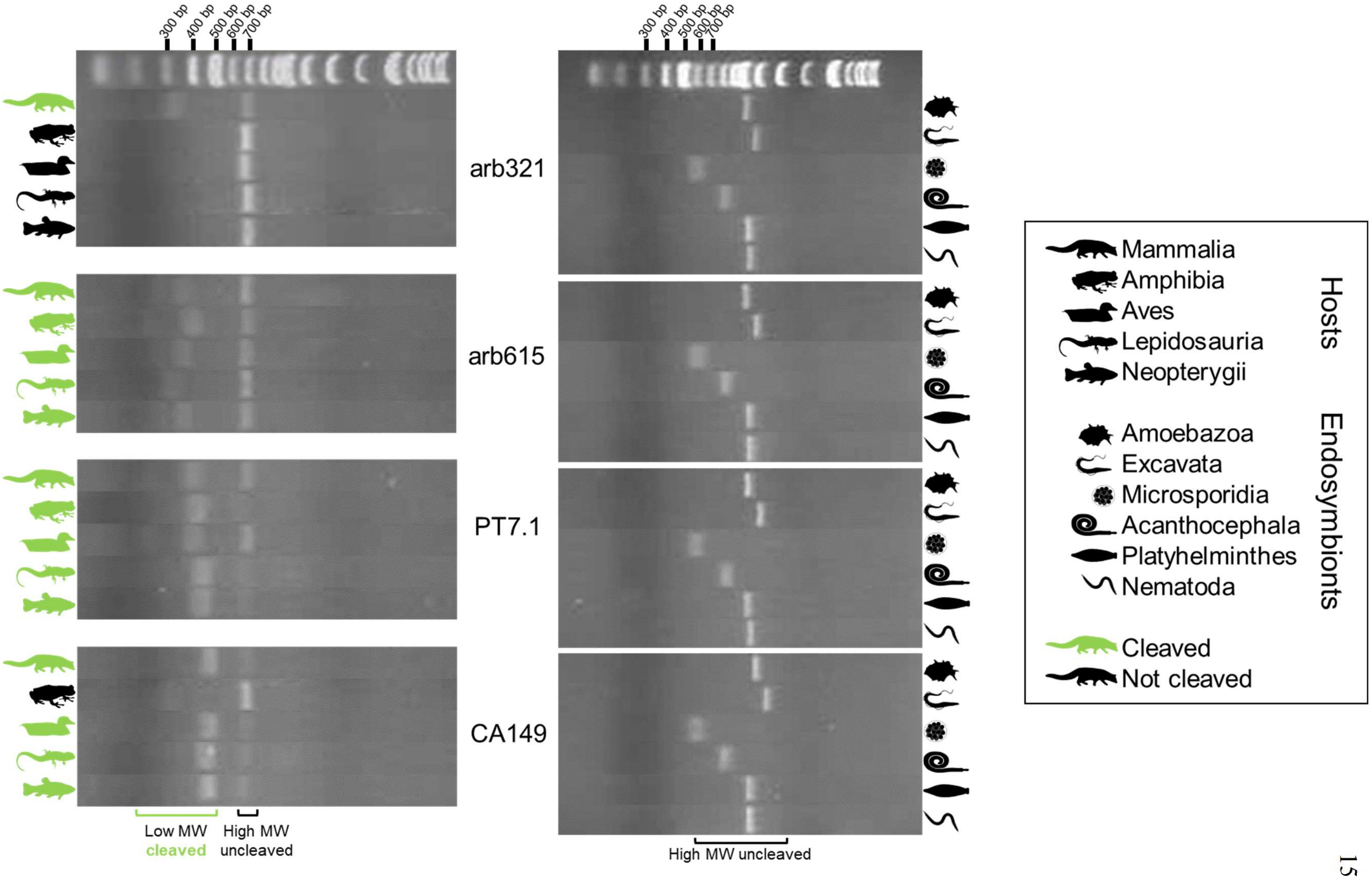

300 bp
400 bp
500 bp
600 bp
700 bp
300 bp
400 bp
500 bp
600 bp
700 bp
arb321
arb615
PT7.1
CA149
Low MW cleaved
High MW uncleaved
High MW uncleaved
Mammalia
Amphibia
Aves
Lepidosauria
Neopterygii
Hosts
Amoebazoa
Excavata
Microsporidia
Acanthocephala
Platyhelminthes
Nematoda
Endosymbionts
Cleaved
Not cleaved

Figure 4. ***In vitro*** **CRISPR-Cas9 digests of host and eukaryotic endosymbiont 18S V4 amplicons.** Gel electrophoresis images show CRISPR-Cas9 digestion products of 18S V4 DNA amplified from vertebrate hosts (**left panel**) and eukaryotic endosymbiotic organisms (**right panel**) with the name of the guideRNA in the center. Sources of substrate DNA are shown as organism icons. Black icons represent organisms not cleaved by CRISPR-Cas9 digest with the specified guideRNA, and green icons represent organisms cleaved by CRISPR-Cas9 with the specified guideRNA. Organisms used for digest were: Mammalia- *Ursus maritimus* (polar bear), Amphibia- *Lithobates chiricahuensis* (leopard frog), Aves- *Gallus gallus* (chicken), Lepidosauria- *Varanus varius* (monitor lizard), Neopterygii- *Salmo trutta* (brown trout), Amoebazoa- *Entamoeba histolytica,* Excavata- *Trypanosoma brucei,* Microsporidia- *Encephalitozoon cuniculi,* Acanthocephala- *Echinorhynchus salmonis,* Platyhelminthes- *Schistosoma mansoni,* Nematoda- *Ascaris suum.* Topmost row is a DNA size standard. Note that 18S V4 amplicon length is variable among eukaryotic endosymbionts and that no eukaryotic endosymbiont amplicons were digested using any of the guideRNAs tested.

Figure 5.

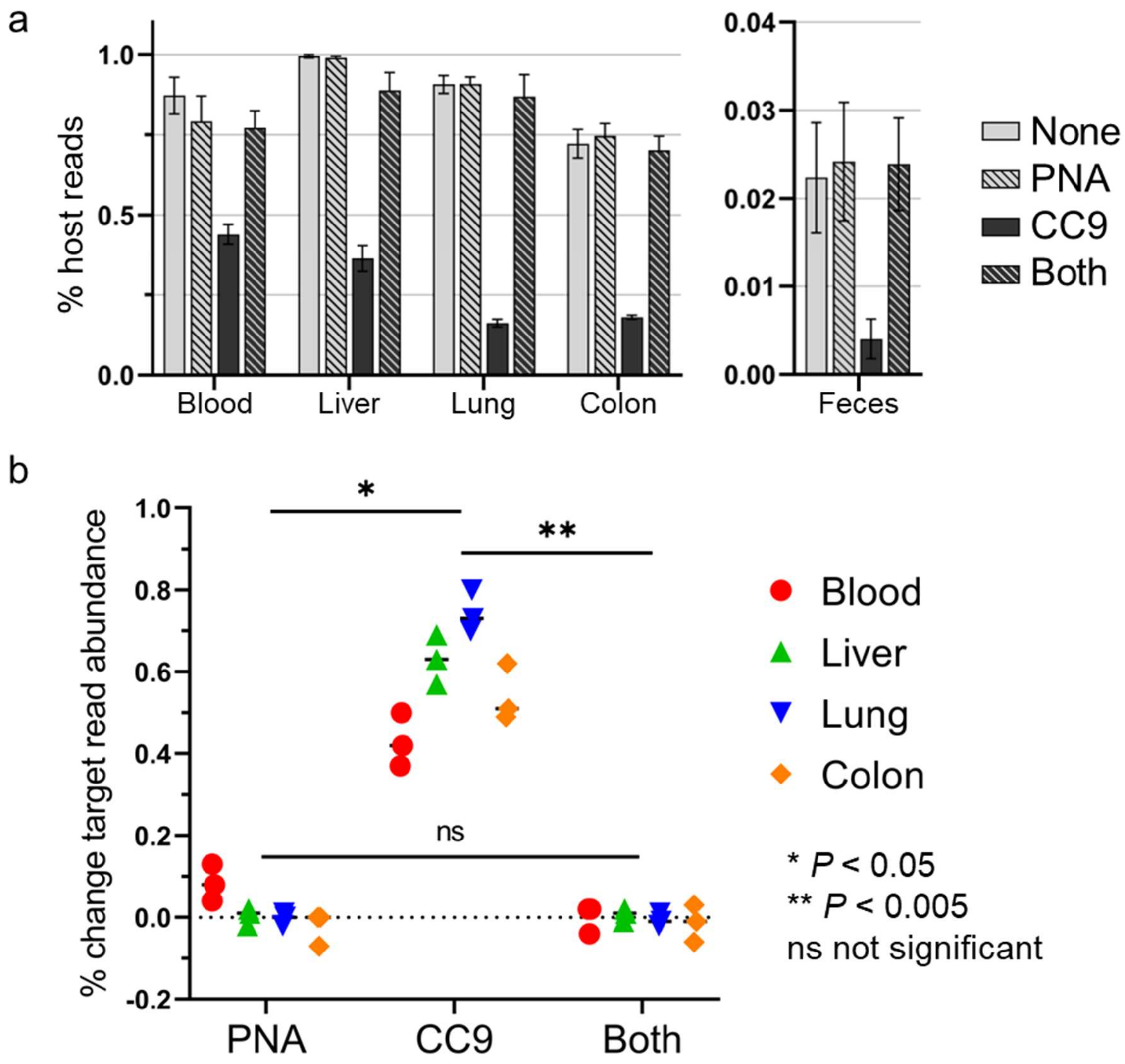


Figure 5. Methods comparison: Host signal reduction with mammal blocking PNA oligo compared to CRISPR-Cas9 amplicon digest in 18S V4 metabarcoding. a, Percent abundance of host reads after quality filtering for five DNA samples metabarcoded under four conditions (triplicate mean): no host signal reduction used (None), published mammal-blocking PNA oligo added to amplicon PCR (PNA), CRISPR-Cas9 digest of amplicons (CC9), and mammal-blocking PNA oligo added to amplicon PCR plus subsequent CRISPR-Cas9 digest of amplicons

(Both). Note scale difference in tissues versus fecal sample. **b,** Results from **a** displayed as percent change in target (non-host) read abundance as compared to no-treatment control for all non-fecal samples. PNA, published mammal-blocking PNA oligo added to amplicon PCR; CC9, CRISPR-Cas9 digest of amplicons; Both, mammal-blocking PNA oligo added to amplicon PCR plus subsequent CRISPR-Cas9 digest of amplicons. CC9 treatment is significantly different from PNA (paired t-test: t = 6.94, df = 3, *P* = .0061) and Both (paired t-test: t = 8.89, df = 3, *P* = 0.0030). See **Supp Table 2** for source data.

Figure 6.

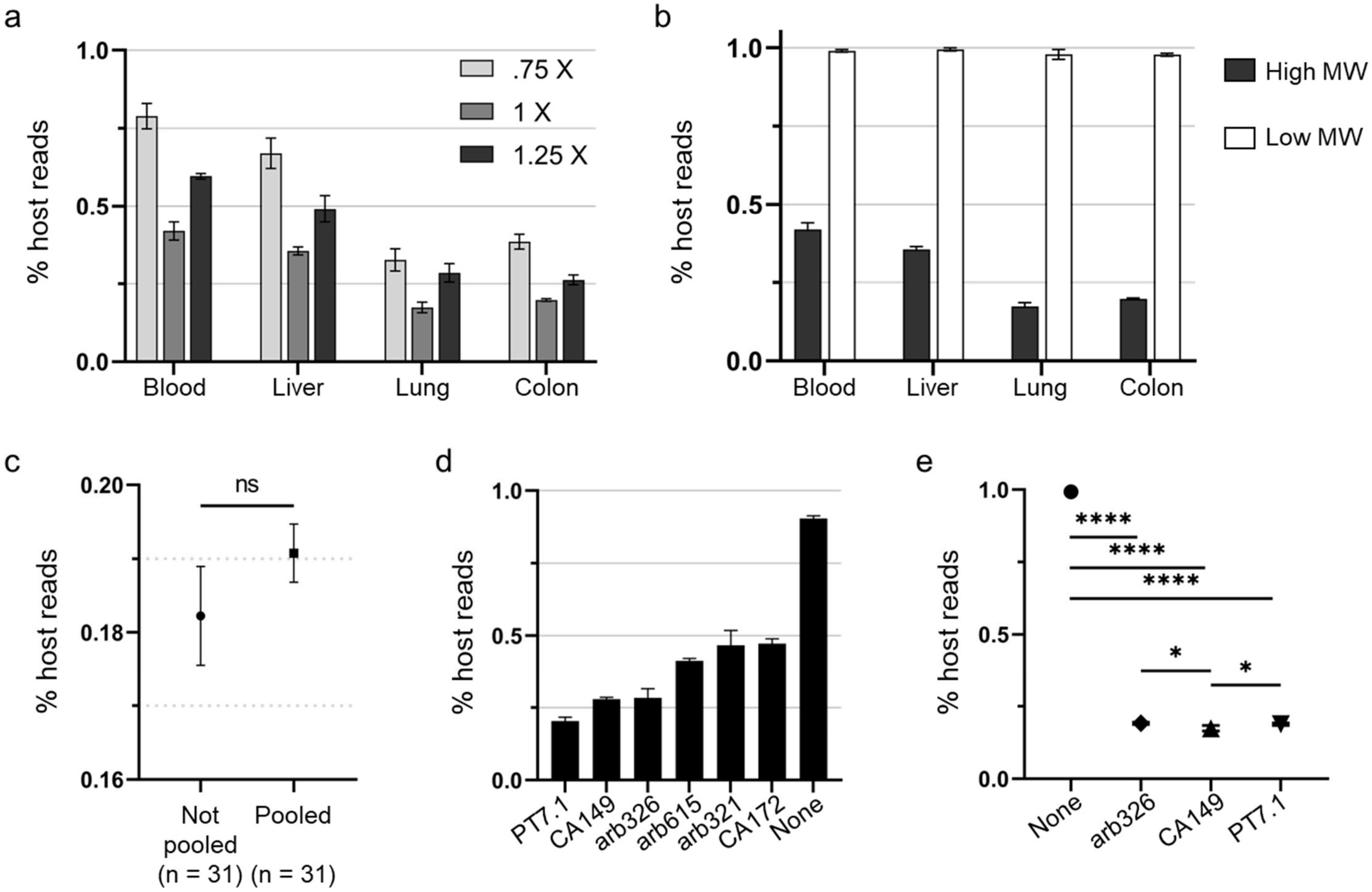

a
% host reads
1.0
0.5
0.0
.75 X
1 X
1.25 X
Blood
Liver
Lung
Colon
b
% host reads
1.0
0.5
0.0
High MW
Low MW
Blood
Liver
Lung
Colon
c
% host reads
0.20
0.18
0.16
ns
Not pooled (n = 31)
Pooled (n = 31)
d
% host reads
1.0
0.5
0.0
PT7.1
CA149
arb326
arb615
arb321
CA172
None
e
% host reads
1.0
0.5
0.0
****
****
****
*
*
None
arb326
CA149
PT7.1

Figure 6. Characterization and optimization of CRISPR-Cas9 mediated host signal reduction in 18S V4 metabarcoding. a, CRISPR-Cas9 (CC9) reaction optimization. Percent host read abundance (triplicate mean +/- SEM) after quality filtering using varying ribonucleoprotein complex (RNP) to DNA target sequence ratios, where 1X represents a 1:1 ratio. **b,** Identity of high and low molecular weight (MW) CC9 cleavage products. Percent host read abundance (triplicate mean +/- SEM) after quality filtering is shown for high and low MW bands extracted after separation by gel electrophoresis. **c,** Comparison of CC9 digest before and after indexing PCR. Mean percent host read abundance +/- SEM after quality filtering is shown for CC9 digest applied to each amplicon prior to library preparation (Not pooled) or to a single pool of amplicons after library preparation (Pooled). ns, not significant (paired t-test: t = 1.38, df = 30, *P* = 0.18). **d,** Effect of short guide RNA (sgRNA) sequence on blood sample 18S V4 metabarcoding. Percent host read abundance (triplicate mean +/- SEM) after quality filtering is shown for 18S V4 amplicons that were not treated with any host signal reduction method (None) or digested with CRISPR-Cas9 using the specified sgRNA prior to library preparation. See Supp Table 3 for source data. **e,** Comparison of sgRNAs in blood sample metabarcoding. Mean percent host reads abundance +/- SEM after quality filtering is shown for three guideRNAs compared to no digest control. * *P* < 0.05, **** *P* < 0.0001, all comparisons not shown are insignificant (paired t-test, df = 30 in all comparisons). See **Supp Table 3** for source data.

Figure 7.

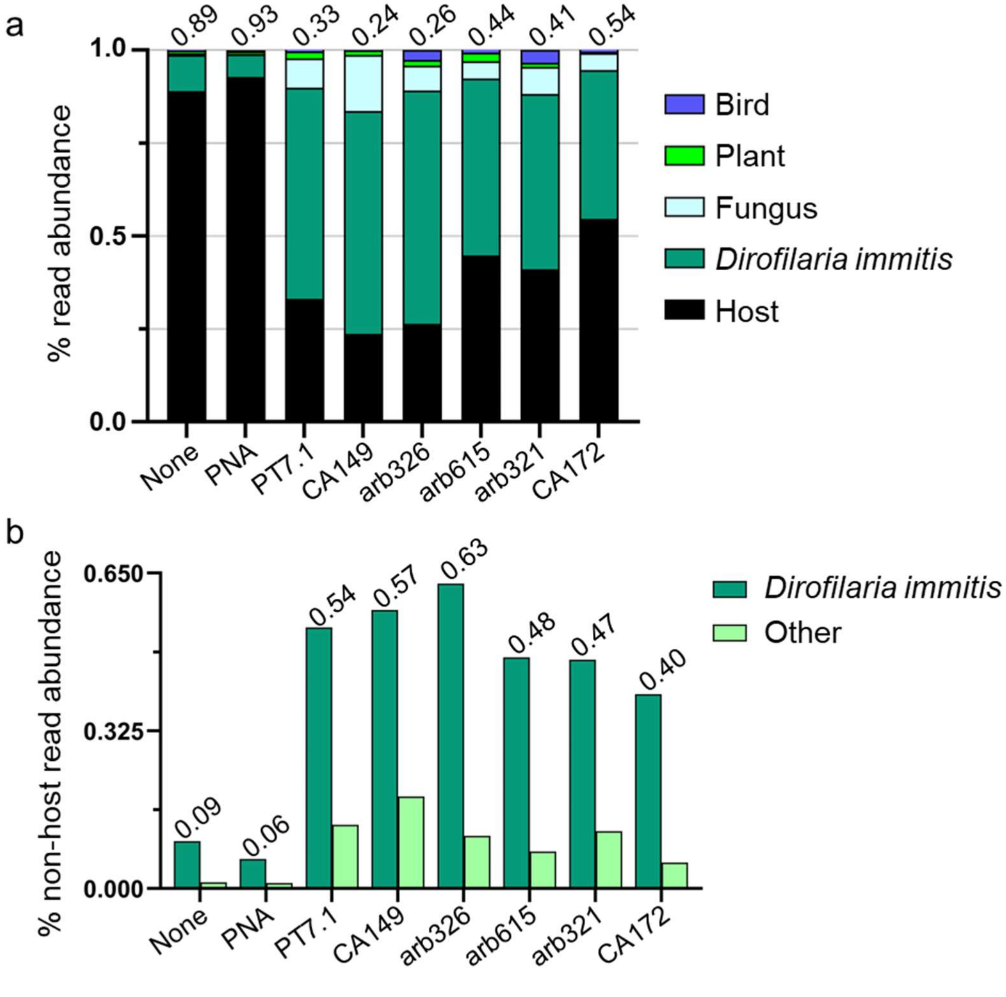


Figure 7. Effect of host signal reduction method on detection of a known parasite infection. Dog blood infected with *Dirofilaria immitis* microfilariae was used as starting material for DNA extraction and 18S metabarcoding. Amplicons were untreated for host signal reduction (None), amplified with a PNA mammal blocker (PNA), or digested with CRISPR-Cas9 using the specified short guide RNAs (X axis). Percent abundance after quality filtering is shown for **a,** all filtered reads or **b,** reads after removing host sequences. Numbers above bars represent **a,** total percentage host reads or **b,** total percentage *D. immitis* reads. Note difference in scale between **a** and **b**. See **Supp Table 4** for source data.

Figure 8.

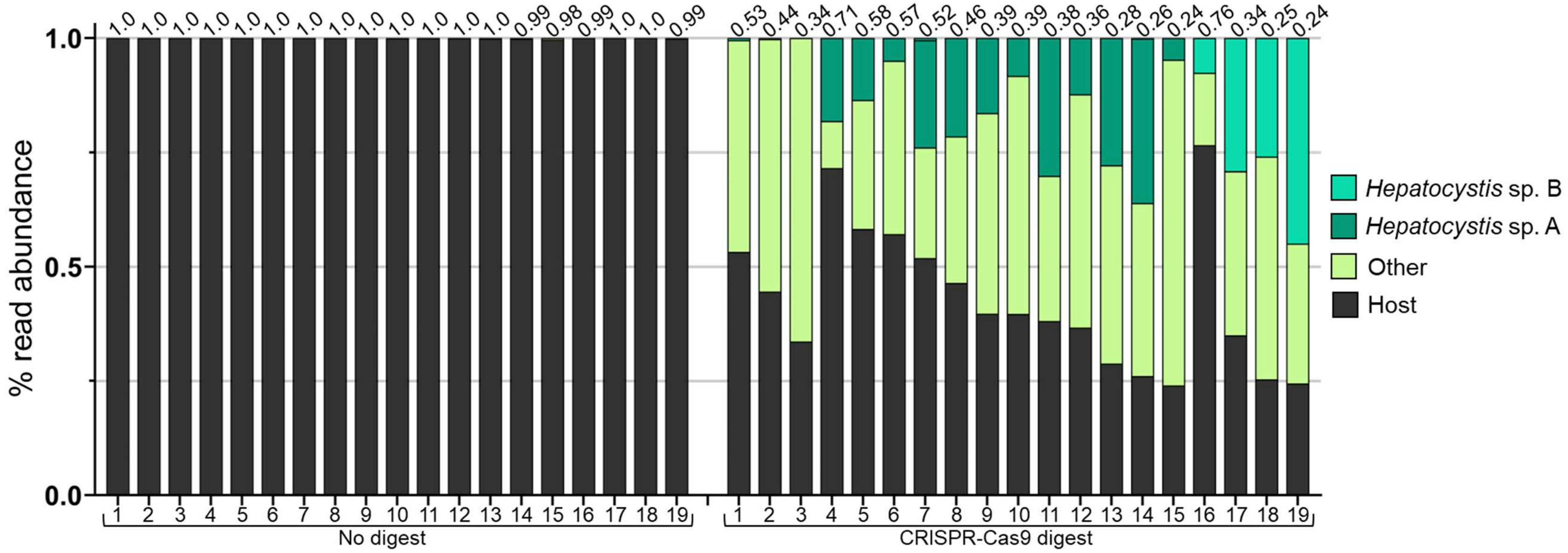


Figure 8. Effect of CRISPR-Cas9 host signal reduction on detection of hemoparasite infection in wild non-human primate blood samples. a, Metabarcoding data are shown as percent read abundance after quality filtering for undigested **(left panel**) and CRISPR-Cas9 digested (**right panel**) amplicons using 19 samples. Reads are categorized as host, *Hepatocystis* spp., and all other reads (Other). Numbers above bars represent total % host reads per sample. No *Hepatocystis* spp. positives were detected by metabarcoding in undigested samples. See **Supp Table 5** for source data.

Supplementary Tables

Supplementary Table 1. Descriptive statistics of read data obtained from 18S V4 metabarcoding with PNA mammal blocker applied to nonhuman primate fecal, blood, and tissue samples.

		% host reads after quality filtering			
Sample	n	Mean	SEM	Min	Max
Feces	6	0.0078	0.0042	0.0000	0.0290
Blood	2	0.8954	0.0644	0.8044	0.9864
Plasma	2	0.7865	0.0410	0.7286	0.8445
Serum	10	0.9018	0.0297	0.6806	0.9928
Brain	2	0.9912	0.0056	0.9832	0.9992
Liver	2	0.9998	0.0000	0.9998	0.9999
Lung	2	0.9732	0.0074	0.9627	0.9837
Spleen	2	0.9999	0.0000	0.9999	0.9999

Supplementary Table 2. 18S V4 region comparison and percent DNA identity between sgRNAs, mouse, and *Trichinella pseudospiralis* sequences.

sgRNA	Target/sgRNA Seq	Mouse Seq NR_003278*	Mouse % ID	Trichinella Seq AY851258**	TP % ID
arb321	AACTGAGGCCATGA**TTAAGA**	AACTGAGGCCATGA**TTAAGA**	100 %	ACCGGAGATAAGT**ATT**G**A**A**A**	55 %
arb326	AGGCCATGATTAA**GAGGGA**	AGGCCATGATTAA**GAGGGA**	100 %	AGATAAGTATTGAA**AGG**A**A**	58 %
arb615	GCAGCTAGGAATAA**TGGAAT**	GCAGCTAGGAATAA**TGGAAT**	100 %	GGTGCATGGAATAATA**GAAT**	75 %
PT7.1	ATTCTTGGACCGGC**GCAAGA**	ATTCTTGGACCGGC**GCAAGA**	100 %	ATTCTTGGATCGCA**GCAAGA**	85 %
CA149	CTCAGCTAAGAGC**ATCGAGG**	CTCAGCTAAGAGC**ATCGAGG**	100 %	NA	0 %
CA172	TCTTAGCTGAGTGT**CCCGCG**	TCTTAGCTGAGTGT**CCCGCG**	100 %	NA	0 %
		Mean	100 %	Mean	45.5 %

sgRNA seed sequences are in bold font; sgRNA sequences and matching bases are highlighted; darker highlighting indicates an overlapping area of two sequences of the same color; TP, *Trichinella pseudospiralis*.

*>NR_003278.3 *Mus musculus* 18S V4 (642 bp)

```
CAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGCTGCAGTTAAAAAGCTCGTAGTTGGATCTTGGGAGCGGGCGGGCGGTCC
GCCGCGAGGCGAGTCACCGCCCGTCCCCGCCCCTTGCCTCTCGGCGCCCCCTCGATGCTCTTAGCTGAGTGTCCCGCGGGGCCCGAAGCGTTTAC
TTTGAAAAAATTAGAGTGTTCAAAGCAGGCCCGAGCCGCCTGGATACCGCAGCTAGGAATAATGGAATAGGACCGCGGTTCTATTTTGTTGGTTT
TCGGAACTGAGGCCATGATTAAGAGGGACGGCCGGGGGCATTCGTATTGCGCCGCTAGAGGTGAAATTCTTGGACCGGCGCAAGACGGACCAG
AGCGAAAGCATTTGCCAAGAATGTTTTCATTAATCAAGAACGAAAGTCGGAGGTTCGAAGACGATCAGATACCGTCGTAGTTCCGACCATAAAC
GATGCCGACTGGCGATGCGGCGGCGTTATTCCCATGACCCGCCGGGCAGCTTCCGGGAAACCAAAGTCTTTGGGTTCCGGGGGGAGTATGGTTG
CAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGGCCTGCGGCTTAATTTGACTCAACACGGG
```

**>AY851258.1 *Trichinella pseudospiralis* 18S V4 (638 bp)

```
CAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGCTGCGGTTAAAACGCTCGTAGTTGAATTGTGGTCTTAGACAACAGTCCCC
CTATATAGGTGTGGCACGGTTGCTCGAGATCTTCATTCGTGGTTGCTGTTGTTGCTCTTCATTGAGTGTCAATGGTGCCTCGAGATTTTACTTTGA
AAAAATTAGAGTGCTCAAAGCAGGTTGTGATGCCTGAATAATGGTGCATGGAATAATAGAATACGATCTCGGTTCTATTTTGTTGGTTTTCGAAC
CGGAGATAAGTATTGAAAGGAACAGACGGGGGCATTCGTATTGCTGCGTTAGAGGTGAAATTCTTGGATCGCAGCAAGATGAACAATTGCGAA
AGCATTTGCCAAGAATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAGGCGATCAGATACCGCCCTAGTTCTAACGGTAAACTATGCC
AACCAGCGATTCGCCGAAGTTCATTTAAGACTCGGCGAGCAGCTTCCGGGAAACCAAAGTGTTTCGGTTCCGGGGGAAGTATGGTTGCAAAGCT
GAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGG
```

Supplementary Table 3. Descriptive statistics of read data obtained from 18S V4 metabarcoding with CRISPR-Cas9 digest, PNA mammal blocker, both CRISPR-Cas9 digest and PNA mammal blocker, or no treatment applied to nonhuman primate blood and tissue samples.

			% host reads after quality filtering			
Sample	Treatment	n	Mean	SEM	Min	Max
Blood	None	3	0.8719	0.0470	0.8044	0.9864
	PNA	3	0.7918	0.0644	0.6983	0.9485
	CC9	3	0.4399	0.0249	0.3819	0.4851
	Both	3	0.8714	0.0413	0.8006	0.9693
Liver	None	3	0.9946	0.0043	0.9841	0.9999
	PNA	3	0.9894	0.0044	0.9806	0.9992
	CC9	3	0.3643	0.0327	0.2924	0.4306
	Both	3	0.9846	0.0032	0.9775	0.9909
Lung	None	3	0.9067	0.0230	0.8736	0.9627
	PNA	3	0.9081	0.0182	0.8746	0.9504
	CC9	3	0.1620	0.0096	0.1406	0.1811
	Both	3	0.9116	0.0246	0.8669	0.9689
Colon	None	3	0.7223	0.0363	0.6752	0.8112
	PNA	3	0.7463	0.0314	0.6769	0.8099
	CC9	3	0.1805	0.0052	0.1688	0.1907
	Both	3	0.7353	0.0210	0.6926	0.7814
Fecal	None	3	0.0224	0.0051	0.0143	0.0347
	PNA	3	0.0242	0.0055	0.0166	0.0376
	CC9	3	0.0040	0.0018	0.0007	0.0083
	Both	3	0.0239	0.0043	0.0162	0.0339

Supplementary Table 4. Descriptive statistics of read data obtained from 18S V4 metabarcoding with CRISPR-Cas9 digest or no treatment applied to nonhuman primate blood samples.

		% host reads after quality filtering			
sgRNA	n	Mean	SEM	Min	Max
PT7.1	3	0.2040	0.0095	0.1905	0.2174
CA149	3	0.2792	0.0046	0.2726	0.2857
arb326	3	0.2829	0.0233	0.2500	0.3158
arb615	3	0.4129	0.0059	0.4045	0.4212
arb321	3	0.4663	0.0366	0.4146	0.5180
CA172	3	0.4718	0.0122	0.4545	0.4891
NONE	3	0.9037	0.0070	0.8937	0.9136
Arb326	31	0.1917	0.0038	0.0974	0.2433
CA149	31	0.1744	0.0080	0.0587	0.2211
PT7.1	31	0.1886	0.0043	0.1027	0.2568
NONE	31	0.9926	0.0033	0.9097	1.0000

Supplementary Table 5. Descriptive statistics of read data obtained from 18S V4 metabarcoding with CRISPR-Cas9 digest or no treatment applied to dog blood samples with known *Dirofilaria immitis* infection.

	% reads after quality filtering				
Treatment	Host	*Dirofilaria immitis*	Fungi	Plant	Bird
None	0.8896	0.0977	0.0037	0.0074	0.0016
PNA	0.9277	0.0608	0.0010	0.0065	0.0039
PT7.1	0.3308	0.5684	0.0780	0.0193	0.0036
CA149	0.2366	0.5999	0.1505	0.0128	0.0002
arb326	0.2634	0.6281	0.0664	0.0155	0.0266
arb615	0.4473	0.4763	0.0461	0.0233	0.0070
arb321	0.4104	0.4716	0.0725	0.0114	0.0339
CA172	0.5460	0.4006	0.0455	0.0030	0.0049
CC9 Mean	0.3724	0.5241	0.0765	0.0142	0.0127

Supplementary Table 6. Descriptive statistics of read data obtained from 18S V4 metabarcoding with CRISPR-Cas9 digest or no treatment applied to red colobus blood samples.

	No treatment				**CRISPR-Cas9 digest**				**No treatment vs CC9**
	% host	% other	% *Hepatocystis* sp. A	% *Hepatocystis* sp. B	% host	% other	% *Hepatocystis* sp. A	% *Hepatocystis* sp. B	% change host reads
1	1.0000	0.0000	0.0000	0.0000	0.5322	0.4634	0.0022	0.0000	0.4678
2	1.0000	0.0000	0.0000	0.0000	0.4455	0.5522	0.0000	0.0000	0.5545
3	1.0000	0.0000	0.0000	0.0000	0.3363	0.6637	0.0000	0.0000	0.6637
4	1.0000	0.0000	0.0000	0.0000	0.7154	0.1028	0.1818	0.0000	0.2846
5	1.0000	0.0000	0.0000	0.0000	0.5820	0.2827	0.1353	0.0000	0.4180
6	1.0000	0.0000	0.0000	0.0000	0.5712	0.3793	0.0494	0.0000	0.4288
7	1.0000	0.0000	0.0000	0.0000	0.5185	0.2420	0.2349	0.0046	0.4815
8	0.9999	0.0001	0.0000	0.0000	0.4639	0.3209	0.2152	0.0000	0.5360
9	0.9999	0.0001	0.0000	0.0000	0.3971	0.4385	0.1643	0.0000	0.6028
10	0.9996	0.0004	0.0000	0.0000	0.3966	0.5205	0.0829	0.0000	0.6030
11	0.9994	0.0006	0.0000	0.0000	0.3810	0.3174	0.3017	0.0000	0.6184
12	0.9994	0.0006	0.0000	0.0000	0.3668	0.5100	0.1232	0.0000	0.6326
13	0.9985	0.0015	0.0000	0.0000	0.2880	0.4337	0.2783	0.0000	0.7105
14	0.9981	0.0019	0.0000	0.0000	0.2604	0.3788	0.3599	0.0000	0.7377
15	0.9966	0.0034	0.0000	0.0000	0.2402	0.7125	0.0473	0.0000	0.7564
16	1.0000	0.0000	0.0000	0.0000	0.7654	0.1586	0.0000	0.0759	0.2346
17	1.0000	0.0000	0.0000	0.0000	0.3494	0.3589	0.0000	0.2912	0.6506
18	0.9998	0.0002	0.0000	0.0000	0.2534	0.4870	0.0000	0.2595	0.7464
19	0.9988	0.0012	0.0000	0.0000	0.2445	0.3058	0.0000	0.4497	0.7543

Supplementary Table 7. Percentage agreement statistics using Cohen's Kappa test of read data obtained from 18S V4 metabarcoding with CRISPR-Cas9 digest or no treatment applied to red colobus blood samples.

		PCR	
		Negative	Positive
Metabarcoding, No HSR	Negative	3	16
	Positive	0	0

# observed agreements	3
% observed agreements	15.79%
Kappa	0
SE of Kappa	0
95% CI	0 to 0

		PCR	
		Negative	Positive
Metabarcoding, CC9 digest	Negative	2	0
	Positive	1	16

# observed agreements	18
% observed agreements	94.74%
Kappa	0.855
SE of Kappa	0.14
95% CI	0.581 to 1.000

CHAPTER 4: **VESPA: an optimized pipeline for metabarcoding-based characterization of vertebrate eukaryotic endosymbiont and parasite assemblages**

Leah A. Owens, Sagan R. Friant, Bruno Martorelli Di Genova, Laura J. Knoll, Monica Contreras, Oscar Noya, Maria G. Dominguez-Bello, and Tony L. Goldber

Abstract

Standardized metabarcoding protocols for bacteria, archaea, and fungi have revolutionized our understanding of microbial communities. Unfortunately, comparable methods for eukaryotic microbiota have lagged due to technical challenges. Despite 54 published studies and several promising results, issues remain with primer complementarity, off-target amplification, and lack of external validation. Here, we present VESPA (Vertebrate Eukaryotic endoSymbiont and Parasite Analysis): a novel, optimized, and validated metabarcoding assay for host-associated eukaryotic community analysis. Using *in silico* prediction, panel PCR, and an engineered mock community standard, we demonstrate VESPA to be more effective at resolving eukaryotic assemblages than previously published methods. When applied to clinical samples from humans and non-human primates, VESPA enabled reconstruction of host-associated eukaryotic endosymbiont communities more accurately and at finer taxonomic resolution than microscopy. VESPA has the potential to advance basic and translational science on vertebrate eukaryotic endosymbiont communities similar to achievements made for bacterial, archaeal, and fungal microbiomes.

Main

Microbiomes are multikingdom assemblages of microorganisms and their entire "theater of activity" including signaling molecules and metabolites[1]. Such communities have emergent properties arising from cross-species and cross-kingdom interactions[2]. One of the most salient examples is the human gut, wherein bacterial community dynamics have direct effects on health[3] and can be manipulated to improve disease outcomes in clinical settings[4]. Evidence is mounting that assemblages of host-associated eukaryotes also form communities with important consequences for host health[5], although they are far less studied compared to their bacterial, archaeal, and fungal counterparts[6]. Even terminology to describe host-associated eukaryotes is lacking. "Eukaryotic microbiome/microbiota"[7] does not include host-associated macro-organisms such as helminths, "nemabiome"[8] is limited to nematodes, and "parasites"[9] excludes commensal/beneficial organisms and includes ectoparasites. Herein we use the term "eukaryotic endosymbionts" to refer to both microscopic eukaryotes (microsporidia, protozoa, algal parasites) and macroscopic metazoans (helminths, pentastomes). In this context, we use the prefix "endo" to include endoparasites and commensals, while excluding ectoparasites (mites, ticks, fleas). We exclude fungi because of their fundamentally different life histories[10] and the fact that established methods already exist for assessing the "mycobiome"[11]. However, we include microsporidia, because their life cycles are considered more similar to protozoa than to fungi[12].

Well-established methods exist to study eukaryotic endosymbiotic organisms. Microscopic observation has been an essential tool since van Leeuwenhoek first described *Giardia* in the seventeenth century[13]. Combined with subsequent advances in staining and enrichment techniques, microscopy is still a gold standard method[14], although it requires specialized training[15] and has inherent resolution limits (i.e., some species cannot be distinguished solely based on morphology, a phenomenon known as “cryptic species complexes”[16]). For example, the genus *Entamoeba* contains pathogenic *E. histolytica* and benign *E. dispar* which appear identical under the microscope[17]. More recently developed molecular assays (e.g., PCR and DNA sequencing of amplicons) have enabled finer taxonomic differentiation, including strain-level identification of species complexes[18]. Although extremely useful, such assays usually have high DNA sequence specificity and are therefore not suitable for characterizing diverse assemblages of eukaryotic endosymbionts.

Methods for characterizing bacterial and fungal assemblages are standardized and based on massively parallel sequencing of amplified marker genes, or metagenomic barcoding (henceforth metabarcoding)[19]. For bacteria, the 16S ribosomal RNA (16S rRNA, or just 16S) locus[20] and for fungi, the internal transcribed spacer (ITS) locus[21] are proven targets for metabarcoding. By contrast, “universal” targets and protocols for metabarcoding of eukaryotic endosymbionts are not standardized[6]. For example, some published methods utilize PCR primer sets originally designed for free-living eukaryotic microbes[22-25], some target metazoans only[26,27], while others focus exclusively on helminths[8,28-30] or gut-associated organisms[31-33]. There is also a conspicuous absence of published comparisons to “gold standard” methods such as microscopy[34]. Moreover, no commercially available reagents exist for assessing the accuracy of eukaryotic endosymbiont metabarcoding-based methods. Community standards (mixtures of organisms or their genetic material in known composition and quantity) have been important for standardizing microbiome protocols and are commercially available[35]. Unfortunately, no such standard exists for eukaryotes other than fungi.

Here we present VESPA (Vertebrate Eukaryotic endoSymbiont and Parasite Analysis), a new methodology for eukaryotic endosymbiont metabarcoding that resolves the issues described above. We compare VESPA to published method*s in silico* and using a new community standard comprised of cloned DNA from eukaryotic endosymbiont lineages across the Tree of Life. We then compare our new method to the "gold standard" of microscopy using clinical samples. Our results show that VESPA and our community standard constitute a major advance that should enable "microbiome-like" insights into the structure and function of vertebrate-associated eukaryotic endosymbiont communities.

Results

Here we compile and evaluate published methods for metabarcoding vertebrate-associated eukaryotic endosymbionts and choose a marker gene and region for amplification. We then compare the relevant subset of published methods to a new method of our own design in a progressive series of experiments. We begin with *in silico* PCR, proceed to amplification of single parasite DNA templates, and then conduct metabarcoding using an engineered mock community standard. We finally apply the best-performing protocol to clinical samples from humans and non-human primates and compare results to those obtained with microscopy.

Methods review and new method design

In a literature review consisting of 54 papers that used amplicon sequencing (metabarcoding) to characterize eukaryotic assemblages in vertebrate hosts (**Supp File 1**), we identified eight marker genes, including nt-MD1 (n = 1), 12S (n = 1), 28S (n = 1), mitochondrial 16S (n = 2), mini-exon Tcl DTU (n = 2), CO1 (n = 2), ITS-2 (n = 13), and 18S (n = 37; **Figure 1a**). Of these publications, 25 targeted specific sub-groups (e.g., nematodes or trypanosomes) and 29 used a pan-parasite/commensal approach. Based on the widespread incorporation of small subunit ribosomal RNA 18S gene (18S hereafter) sequences into databases, the standardized use of the counterpart prokaryotic 16S gene for bacterial metabarcoding, and evidence that non-protein coding genes outperform protein-coding genes as metabarcoding markers[36], we chose to pursue 18S as our marker gene.

18S contains hypervariable regions V1 - V9, and the regions most commonly targeted in the studies reviewed were V4 (n = 13) and V9 (n = 13; **Figure 1b**). The 18S V4 region has the highest entropy within the size limits of MiSeq v2 chemistry[37] and therefore the highest taxonomic resolution for this commonly used metabarcoding platform, so we chose to target this region. We identified a total of 22 published sets of V4 primers. Additionally, we created new 18S V4 primers designed to target all eukaryotic endosymbionts, consisting of 4 candidate forward primers and one reverse primer (see methods section for details on primer design, **Supp Info Table 1** for primer sequences, and **Figure 1c** for a map of primer binding sites).

Testing metabarcoding methods for taxonomic coverage using *in silico* PCR

Testing all 22 published 18S V4 primer sets *in silico* yielded an average eukaryotic endosymbiont coverage of 64.9 % (**Table 1,** bolded columns). No primer set recognized both *Plasmodium* and *Giardia*, and 9 of 19 did not recognize either (**Table 1,** final two columns). We found significant off-target coverage (> 5 %) of bacterial and/or archaeal groups for 4 of 22 sets (**Table 1,** asterisks), and the primer set with the highest overall eukaryotic coverage (96.3 %; Hugerth 2014 "563/1132”) also had the highest coverage of archaea and bacteria (47.9 % and 72.0 % respectively; **Table 1**). Primer sets with > 5 % off-target coverage were not analyzed further.

Table 1. ***In silico* taxonomic coverage for published 18S V4 primer sets**

		Off-target groups		**Eukaryotic endosymbiont groups**		Specific examples	
	n =	20,197	381,535	4,229	15,265	198	23
Primer set ID	Primers	Archaea	Bacteria	**Helminths**	**Protozoa**	*Plasmodium*	*Giardia*
Bates[38]	515f/1119r	0	0	**80.4**	**95.9**	94.8	0
Bower*[39]	18SEUK581F/18SEUK1134R	46.2*	8.2*	**0.4**	**82.4**	0	72.7
Bradley[37]	TAReuk454FWD1/V4r	0	0	**48.9**	**67.1**	97.9	0
C-S[40]/Brate 2[41]	3NDf/V4_euk_R2	0	0	**50.8**	**22.8**	0	0
C-S[40]/Brate 1[41]	3NDf/V4_euk_R1	0	0	**5.8**	**21.1**	0	0
C-S[40]/Geisen[42]	3NDf/1132mod	0.3	0	**80.7**	**94.2**	0	0
Comeau[43]	E572F/E1009R	0	0	**65.3**	**44.5**	0	0
DeMone**[44]	18SV4_F/_R/Giardia_R	0	0	**86.4**	**62.3**	42.8	0
Hadziavdic 566[45]	F-566/R-1200	0	0	**76.4**	**81**	99.6	0
Hadziavdic 574[45]	F-574/R-952	0	0	**48.3**	**62.9**	61.3	0
Hugerth 574*[46]	574/1132	12.5*	0	**80**	**94.2**	0	0
Hugerth 616[46]	616/1132	3.3	0.2	**93.1**	**75.8**	0	45.5
Hugerth 563*[46]	563/1132	47.9*	72*	**96.1**	**96.4**	0	100
Krogsgaard**[32]	G3F1/R1/G4F3/R3/G6F1/R1	0	0	**78.5**	**67**	94.8	0
Machida[47]	18S#1/18S#2RC	0	0	**78.1**	**45.2**	97.9	0
Sikder*[48]	MMSF/MMSR	17.5*	0	**79.3**	**42.7**	0	0
Stoeck[49]	TAReuk454F1/R3	0	0	**49.1**	**78.4**	97.9	0
Wood[50]	Nem18SlongF/Nem18SlongR	0	0	**32.2**	**25.2**	2.6	0
Zhan[51]	Uni18S/R	0	0	**72.8**	**64**	0	100

Numbers shown are % coverage allowing for 1 mismatch with a 2-base pair 3' window using the SILVA 138.1 SSU rRNA NR Ref database; n, number of total eligible accessions; * removed from further analysis due to high prokaryotic complementarity; ** multiple primer sets were combined for analysis. See **Supp Info Table 1** for full primer names and sequences.

In silico PCR including our 4 new primer sets alongside the remaining 18 published 18S V4 sets yielded coverage data spanning a wide range (5.8 % to 98.0 %; **Table 2**). Across target groups (normalizing by eligible accessions), our newly designed primers had the highest mean percent coverage, at 95.2 % - 96.8 %, and also the best complementarity as evidenced by the lowest score in a rank sum analysis (**Table 2**, final column).

Table 2. ***In silico*** **taxonomic coverage of helminths and protozoa for published and newly designed 18S V4 primer sets.**

		n = 3,097	n = 2,913	Rank		**Rank sum**
Primer set ID	**Mean**	Helminths	Protozoa	Helminths	Protozoa	
Owens 29F	**96.8%**	95.5%	98.0%	1	1	**2**
Owens 2-2b	**96.4%**	94.9%	97.9%	2	2	**4**
Owens 13F	**96.4%**	94.9%	97.9%	2	2	**4**
Owens 9F	**95.2%**	94.4%	96.0%	4	4	**8**
Bates	**88.2%**	80.4%	95.9%	8	5	**13**
Hugerth	**84.5%**	93.1%	75.8%	5	8	**13**
Krogsgaard**	**81.0%**	93.0%	69.0%	6	9	**15**
Hadziavdic 566	**78.7%**	76.4%	81.0%	10	6	**16**
DeMone**	**75.3%**	86.4%	64.1%	7	11	**18**
Stoeck	**63.8%**	49.1%	78.4%	13	7	**20**
Machida	**61.7%**	78.1%	45.2%	9	14	**23**
Bradley	**58.0%**	48.9%	67.1%	14	10	**24**
Hadziavdic 574	**55.6%**	48.3%	62.9%	15	12	**27**
Comeau	**54.9%**	65.3%	44.5%	11	15	**26**
C-S/Geisen*	**47.5%**	40.7%	54.2%	16	13	**29**
C-S/Brate 2*	**36.8%**	50.8%	22.8%	12	17	**29**
Wood*	**28.7%**	32.2%	25.2%	17	16	**33**
Zhan*	**14.0%**	5.7%	22.3%	19	18	**37**
C-S/Brate 1*	**13.5%**	5.8%	21.1%	18	19	**37**

Shaded rows, primers designed in this study; %, % coverage calculated allowing for 1 mismatch with a 2-base pair 3' window using the SILVA 138 SSU rRNA NR Ref database; n, number of eligible accessions; Mean, mean coverage of all parasite/commensal groups; * < 50 % overall mean target complementarity; ** multiple primer sets combined for analysis. See **Supp Info Table 1** for full primer names and sequences.

In silico coverage analysis using finer-resolution groups (**Figure 1d**) showed that our new primers consistently amplified (defined as coverage of 50 % or higher) all 24 clades of eukaryotes tested whereas no other primer sets did. Particularly problematic were *Giardia* (recognized by our primers and one other set in which a second reverse primer must be used to specifically amplify *Giardia*), *Microsporidia* (recognized by our primers and two other sets), and *Trichomonadea* (recognized by our primers and three other sets; **Figure 1d,** red boxes).

Testing metabarcoding methods for on-target amplification using purified DNA

In PCR amplification of genomic DNA (gDNA) from 22 individual eukaryotic endosymbiont organisms (**Supp Info Table 2**), all four sets of candidate primers amplified more organisms than did any of the published primer sets (Owens 29F: 22 of 22, Owens 2-2bF: 21 of 22, Owens 13F: 20 of 22, Owens 9F: 20 of 22), followed by the Bates (19 of 22), Hadziavdic 566 (18 of 22), and Stoeck (16 of 22) sets (**Figure 1e**). Furthermore, two of the new sets were the only primers to successfully amplify 18S V4 from *Giardia* gDNA (Owens 29F and Owens 2-2bF), as expected based on *in silico* data (**Figure 1e,** red box).

Testing metabarcoding methods for amplification bias using a community standard

Community standards are not available for eukaryotic endosymbionts, so we collected protozoa (n = 10), helminths (n = 5), and a microsporidian (n = 1) (**Supp Info Table 3**) from various sources (e.g., specimen repositories, veterinary post-mortem examinations). We then isolated 18S genes from these samples and mixed them at an equimolar ratio to create a community standard, which we named "EukMix" (**Figure 2a**). Metabarcoding EukMix as input with previously published and newly designed primers allowed us to directly compare empirical read abundances for each organism to their predicted (equal) abundances (**Figure 2b**). The abundances of six organisms were underestimated by every primer set, and the abundances of three organisms were overestimated by every primer set (**Supp Info Table 4**), but the absolute mean difference from theoretically equal abundance was lowest with newly designed primer set

Owens 29F (**Figure 2c**), which also yielded abundance data statistically significantly closer to actual input levels than any other set tested (**Figure 2d**). Primer set Owens 29F consistently reconstructed the EukMix community most accurately (i.e., evenly), as determined by standard diversity/evenness measures (Pielou's species evenness, Simpson's diversity index, and Shannon diversity; **Figure 2e**). We therefore chose 29F/21b8R as the primer set for our finalized VESPA metabarcoding protocol (see accompanying **Protocol**).

VESPA compared to microscopy

Humans VESPA analysis of 12 human clinical samples yielded high-quality data (**Supp Info Table 5**) including low proportions of off-target prokaryotic reads (**Figure 3a**) and host reads (host read mean = 2.97 % per sample, range: 0.11 % - 17.4 %) and correspondingly high proportions of endosymbiont reads (**Figure 3b, 3c**).

VESPA successfully identified all three helminth and seven protozoan taxa identified with microscopy (**Figure 3d**) and found these taxa in more individuals than did microscopy, with 61.4 % of positive samples identified solely by VESPA (**Figure 3e**). Conversely, no positives were identified by microscopy alone. Four additional taxa were found exclusively by VESPA, including one helminth, *Trichuris trichuria* (1 positive of 12 samples*),* and three protozoa*,* *Entamoeba hartmanni* (10 positives of 12 samples)*,* *Enteromonas hominis* (3 positives of 12 samples)*,* and *Pentatrichomonas hominis* (1 positive of 12 samples). Three of 12 patients were known by taxon-specific PCR to be infected with *Onchocerca,* which is not visible microscopically in feces, and all 3 were positive by VESPA. Overall, taxonomic richness was statistically significantly higher by VESPA than by microscopy for both helminths (mean richness = 0.5 by microscopy, 1.92 by VESPA, Wilcoxon matched-pairs signed rank test, 2-

tailed, *P* = 0.001) and protozoa (mean richness = 2.33 by microscopy, 5.67 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.0005; **Figure 3f,** left panel). Prevalence was also higher by VESPA for helminths (mean prevalence = 0.25 by microscopy, 0.60 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.25) and protozoa (mean prevalence = 0.23 by microscopy, 0.54 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.002; **Figure 3f,** right panel).

Non-human primates VESPA analysis of 40 non-human primate clinical samples yielded high-quality sequencing reads (**Supp Info Table 5**) with low proportions of off-target prokaryotic reads (**Figure 4a**) and host sequence reads (host read mean = 3.2 % per sample, range: 0 % - 18.49 %) and correspondingly high proportions of endosymbiont reads (**Figure 4b, 4c**).

VESPA successfully identified all eight helminth and six protozoan taxa identified with microscopy (**Figure 4d**) and found these taxa in more individuals than did microscopy, with 47.08 % of positive samples identified by VESPA only (**Figure 4e**). One positive out of 29 total for a helminth (*Physaloptera* sp. 1) and 2 positives out of 28 total for a protozoan (*Balantidium coli)* were identified by microscopy only. Six additional taxa were found exclusively by VESPA: *Entamoeba chattoni* (16 positives of 40 samples), *Endolimax nana* (19 positives of 40 samples), *Enteromonas* sp. (6 positives of 40 samples), *Piroplasmida* sp. (2 positives of 40 samples), *Blastocystis* sp. (38 positives of 40 samples), and *Enterocytozoon bieneusi* (3 positives of 40 samples; **Figure 4d, 4e**). *Piroplasmida* are intraerythrocytic parasites not visible in fecal samples and were found in 2 of 40 samples with VESPA. Thirty-one samples were positive for the *Entamoeba histolytica/dispar* species complex by microscopy and the same 31 samples were found to be positive by VESPA but could be further taxonomically resolved as *Entamoeba dispar* in all cases. Richness was higher by VESPA than by microscopy for helminths (mean richness = 1.73 by microscopy, 2.13 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.0009), protozoa (mean richness = 2.8 by microscopy, 5.5 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* < 0.0001), and microsporidia (mean richness = 0 by microscopy, 0.08 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.25; **Figure 4f,** left panel). Prevalence was also higher by VESPA than by microscopy for all three parasite groups (helminth mean prevalence = 0.22 by microscopy, 0.26 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.33; protozoa mean prevalence = 0.22 by microscopy, 0.43 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.002; microsporidia mean prevalence = 0 by microscopy, 0.8 by VESPA, Wilcoxon matched-pairs signed rank test, 2-tailed, *P* = 0.25; **Figure 4f,** right panel).

Discussion

To identify a single method for the “universal” identification of vertebrate-associated eukaryotic endosymbionts in community assemblages, we analyzed published approaches and found a wide range of amplification targets and protocols. From this literature review, we chose to focus on the 18S V4 locus and designed new primers to recognize all known groups of eukaryotic endosymbionts. We then tested published primers and our newly designed primers in a series of experiments *in silico* and *in vitro* to determine which protocols, if any, could accurately reconstruct eukaryotic endosymbiont communities. Our results clearly show that metabarcoding using newly designed primer set 29F recognizes the greatest range of endosymbionts of interest with the least off-target amplification and PCR bias of any method tested. We name our new method VESPA (Vertebrate Eukaryotic endoSymbiont and Parasite Analysis).

VESPA recognized more eukaryotic endosymbiont groups *in silico* than did other published methods tested, including methods that used multiple primer sets to increase coverage. Multiple primer sets, usually involving multiple independent PCR amplifications, are a feasible strategy for increasing coverage[32,33]. However, this approach adds reagent costs and presents technical challenges related to sequencing and bioinformatics[52,53]. Our single primer set approach should therefore reduce barriers to entry for adopting our new method. We then corroborated these *in silico* results with amplification of purified targets and similarly found that our primer sets amplified the greatest range of single organisms *in vitro*.

To examine the performance of published methods and VESPA, we directly compared assays by using an equimolar community standard, EukMix, as input for metabarcoding. Results from VESPA reflected the underlying composition of the community standard more accurately than did results from other assays. The EukMix community standard should be useful for quality control in laboratories choosing to adopt our method, and for standardization and validation, much as community standards containing bacteria and fungi have enabled standardization of microbiome protocols[35,54]. We note that the relationship between sequencing reads and organism abundance or biomass is complicated by wide variation in 18S copy number among eukaryotic endosymbionts[34,55]. Copy number corrections have been applied in studies of other systems[56,57], and such corrections could prove useful for investigations where quantifying organism abundance or biomass are the desired outputs.

Compared to microscopic examination, VESPA detected protozoa, microsporidia, and helminths in more individuals, identified additional organisms, resolved a cryptic species complex, and identified organisms not visible in fecal samples. We suspect that the greater sensitivity of VESPA results from the nature of molecular amplification – namely, that PCR can detect a theoretical minimum of one molecule of target DNA[58]. Microscopy-negative samples that were PCR positive by VESPA may not have contained intact organisms or their eggs or may even have been positive by virtue of the presence of small amounts of cell-free DNA[59]. In this light, we caution that our method will likely be most useful for applications where the presence of eukaryotic endosymbiont DNA is itself taxonomically informative, regardless of whether that DNA represents an intact or viable organism.

Because of the labor-intensive nature of microscopy and its dependence on trained experts, VESPA will also be useful for studies which are large-scale or performed in multiple laboratories, where labor costs and inter-observer variability would otherwise be impractical. In this light, we note that microscopy identified three positive samples not identified by VESPA in non-human primates. We suspect that these findings may represent microscopy false positives, especially because these two taxa (*Physaloptera* and *Balantidium*) are notoriously difficult to identify morphologically[60,61].

Our contribution with this work is a publicly available protocol for metabarcoding eukaryotic endosymbiont communities that outperforms published methods by every measure examined. VESPA is intentionally designed to have broad applicability, from microbial ecology to parasitology to clinical diagnostics. Although we tested VESPA using Illumina sequencing technology, it should be readily adaptable to other amplicon sequencing technologies available now and in the future. VESPA is compatible with existing bacterial and fungal pipelines, with metabarcoding of all three taxa run on the same sequencing platform. Addition of VESPA to established protocols for characterizing bacterial microbiomes and mycobiomes could have far reaching benefits. For example, it has been suggested that studies of the human gut microbiome should routinely incorporate analyses of eukaryotic diversity in order to capture overall microbial community function[5]. VESPA can provide this missing eukaryotic component and thereby enable cross-kingdom characterization of microbial ecosystem structure and function, opening new avenues for basic and applied research.

Methods

Methods review and new method design

Literature searches were performed in January 2021 and updated in January 2023. Search terms or combinations of search terms including “Metagenomics,” “Metagenomic barcoding,” “Metabarcoding,” “Targeted amplicon deep sequencing,” “Eukaryotic microbiome,” “Gastrointestinal,” “Gut,” “Parasite,” and “18S” were used to query PubMed, Web of Science, and Google Scholar. Results were manually evaluated for relevance and details were compiled in an excel spreadsheet (**Supp File 1**). We identified 96 studies including reviews and methods papers, 54 of which were primary research on vertebrate-associated eukaryotes. We chose to focus on 18S because in previous metabarcoding studies, non-coding genes outperformed coding genes[36,62], 18S has islands of conserved sequence interspersed with areas of high entropy (hypervariable regions), allowing broad priming for coverage and diverse amplicons for resolution[45], and database coverage for 18S is higher than for other loci[63]. Of the 9 hypervariable

18S regions, V4 has the highest taxonomic resolution[37], so we focused on this region and identified 22 sets of published V4 primers (**Supp Info Table 1**).

We also designed new 18S V4 primers with the goal of amplifying all eukaryotic endosymbiont groups with little to no prokaryotic complementarity. We began by creating a database of parasite/commensal 18S rRNA sequences containing representatives from all phylogenetic lineages containing at least one vertebrate-associated eukaryotic endosymbiont. We downloaded sequences from all known groups of endoparasites/endosymbionts from NCBI Genbank[64] or the SILVA 138.1 Small Subunit rRNA Non-Redundant Reference Database (n = 510,508 total accessions[63,65]; SILVA Ref NR hereafter) at a depth of one species per genus, beginning with the Centers for Disease Control's "Alphabetical Index of Parasitic Diseases" [66]. To ensure broad coverage of commensals, zoonoses, and novel organisms we added non-pathogenic protozoans of humans[67], parasites/commensals of great apes[68], and parasites of veterinary importance[69]. We then used MUSCLE[70] implemented in MEGA 11[71] to align the resulting 658 full-length 18S sequences, which covered a broad range of pathogenicity, vertebrate hosts, and tissue tropisms. To identify candidate conserved regions, we utilized the Arb software suite[72], and the ecoPrimers function in OBItools[73], with manual inspection and adjustment as needed. We then extracted every 16 - 20-mer candidate sequence within those regions and tested them for taxonomic coverage against SILVA Ref NR using the SILVA TestProbe and TestPrime tools[74]. Candidate primers with high overall complementarity were manually adjusted for maximum coverage.

We aimed to avoid degeneracy as it has been shown to create bias in 18S V4 amplification[37] and succeeded in the forward primer. Degeneracy was required in the reverse primer, although not in the four terminal 3′ nucleotides. Furthermore, of the three degenerate positions in the reverse primer, no targeted groups required all three degeneracies, and most required just one. To increase homogeneity and avoid potential biases against rare sequences, we used 5-deoxyinosine in the four-fold degenerate position instead of N, thereby limiting our reverse primer mixture to four distinct oligonucleotides[75].

The forward region identified for priming had higher GC content than the reverse region, so we forewent the standard guidelines for GC content and melting temperature differences in order to prioritize coverage, with the knowledge that we could later add Locked Nucleic Acids (LNAs) to modify the melting temperature if needed[76]. In the end, this modification was not necessary because the DNA polymerase for PCR (described below) tolerates a wide melting temperature range and has a universal annealing temperature regardless of primer sequence. In total we designed 4 forward primers and one reverse primer (**Supp Info Table 1)** for further testing.

Testing metabarcoding methods for taxonomic coverage using *in silico* PCR

For the initial analysis of published protocols for taxonomic coverage, we used locus-specific sequences (i.e., not including linkers, adapters, or barcode elements) from all 22 18S V4 primer sets identified in our literature search (**Supp File**, **Supp Info Table 1**). *In silico* PCR of SILVA Ref NR was performed using the TestPrime tool allowing for a single mismatch and a mismatch-free two base pair 3′ window. For this analysis, "helminth" accessions included Acanthocephala (n = 66), Nematoda (n = 2,170), and Platyhelminthes (n = 1,993) and "protozoa" accessions included Amoebozoa (n = 1,148), Discoba (n = 1,032), Excavata (n = 389), Alveolata (n = 9,140), and Stramenopiles (n = 3,556). In two cases where multiple primer sets were used in combination (Krosgaard - three sets and DeMone - two sets), we tested each set individually and conservatively estimated coverage by reporting only the highest percentage for each taxon. Primer sets with > 5 % coverage of off-target prokaryote groups (archaea and bacteria) were not analyzed further (n = 4 sets).

In silico PCR was then used to evaluate the published primer sets remaining (n = 18) alongside our new candidate primers (n = 4; **Supp Info Table 1**). At this stage, we filtered target sequences to contain only parasites of vertebrates because the inclusion of environmental/free-living organisms can distort parasite coverage metrics. Specifically, we split clades that contained both free-living organisms and parasites of invertebrate hosts (*e.g., Rhabditida* and *Entamoeba*) into higher-resolution, curated groups. We included free-living, opportunistic parasites of clinical importance, including *Balamuthia mandrillaris* and *Naegleria fowleri*, and we excluded sequences whose label in the SILVA database was incorrect (i.e., the taxonomy string associated with the record did not match the phylogenetic placement in the guide tree; n = 14). Coverage metrics were normalized to eligible accession numbers, which were similar across primer sets because of similar priming locations in the V4 region (see **Figure 1c** for primer map). We compared taxonomic coverage for primer sets using the TestPrime tool[74] and SILVA Ref NR[63,65] allowing for a single mismatch with a mismatch-free two base pair 3′ window. Primers with ≤ 50 % overall mean coverage of target groups and methods that required more than a single primer set were not considered further.

Testing metabarcoding methods for on-target amplification using purified DNA

We assessed amplification success of the remaining 4 newly designed and 8 published primer sets across parasite groups using 22 genomic DNA (gDNA) isolates from single vertebrate endoparasites as template for PCR. Samples were obtained from reputable reagent repositories and expert parasitologists (for sample details including sources see **Supp Info Table 2**) either as purified DNA or whole organisms. gDNA from whole worms and pelleted protozoal cultures were extracted using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) using 0.2 g of starting material, eluted in Qiagen buffer AE, and stored at -20 °C. PCR conditions were as follows: 1 X Platinum II Hot Start PCR MasterMix (ThermoFisher, Waltham, Massachusetts, USA), 0.2 µM forward primer with Nextera adapter, 0.2 µM reverse primer with Nextera adapter, ThermoFisher 0.2 X Platinum II GC Enhancer, 0.8 ng/µl gDNA in a total 12.5 µl reaction; 94 °C for 2 minutes, 30 cycles of [94 °C for 15 seconds, 60 °C for 15 seconds, 68 °C for 15seconds], and hold at 4 °C. Products were electrophoresed on a 1.5 % agarose gel with SYBR gold DNA dye (ThermoFisher) and a 1 kb DNA size standard. Amplification was scored by band presence on an agarose gel upon visualization under UV illumination with a GelDoc XR imager (BioRad, Hercules, California, USA).

Testing metabarcoding methods for amplification bias using a community standard

Preliminary metabarcoding experiments using mixes of gDNA from single parasites demonstrated a non-linear relationship between DNA input and sequence read abundance, likely due to rRNA copy number variation[77]. We addressed this issue by extracting, amplifying, and cloning parasite DNA from 16 vouchered parasite specimens from verified sources or identified by experts (**Supp Info Table 2**). 18S rDNA sequences were amplified with full-length universal or group-specific primers (see **Supp Info Tables 1, 3**) using Qiagen HotStar Plus Taq DNA polymerase according to manufacturer's instructions. Products were verified for size on an agarose gel and Sanger sequenced. Correct 18S sequences were cloned into a pCR4-TOPO vector using a TOPO TA Cloning Kit for Sequencing (Invitrogen, Waltham, Massachusetts, USA) and Invitrogen One Shot competent cells according to manufacturer's instructions. Colonies were screened by PCR and Sanger sequencing. Plasmid DNA (plDNA) extracted from verified transformants was mixed at equimolar ratios to create the equimolar EukMix community standard reagent. This strategy assures equal 18S copy number input among organisms, which, in the case of amplicon sequencing, enables assessment of primer bias and potential of the assays to yield quantitative data[78].

Metabarcoding using new and published primer sets was performed in triplicate with community standard as starting material using the procedure described below. Resulting sequencing reads were filtered for quality using a cutoff of Q = 30 and mapped to a database containing full-length 18S sequences of clones comprising the EukMix mock community using a mapping stringency of 99 % similarity and 99 % length fraction in CLC genomics workbench v.10.2 (Qiagen). The resulting abundances for each community standard component were used to calculate evenness metrics in R v.3.6.3, and GraphPad Prism v.8.4.3 was used for graphing data and for statistical analyses.

VESPA compared to microscopy

Sample collection Clinical samples used in this work were excess material from concluded studies that had been previously evaluated for eukaryotic endosymbionts using microscopy. Human fecal samples had been collected from communities on the southern Venezuelan border with Brazil[79]. Non-human primate fecal samples were collected from semi-free ranging Nigerian red capped mangabeys (*Cercocebus torquatus*) in a sanctuary[80]. Appropriate IRB approvals (IVIC IRB #DIR-0609/1542/2015) and IACUC protocols (The University of Wisconsin-

Madison's IACUC protocol # V1490) were obtained by each collaborator and all samples were completely de-identified prior to use.

Microscopy Microscopic analyses of non-human primate and human feces were performed as previously described[81]. Briefly, one gram of formalin preserved feces was concentrated via formalin-ethyl acetate sedimentation[80] and the sediment was examined in its entirety at ×10 objective light magnification for gastrointestinal parasites by an expert parasitologist. Additionally, one drop of sediment from each sample was examined at ×40 objective light magnification for identification of protozoa.

Genomic DNA isolation Human fecal samples were processed to remove bacteria and debris as previously described[82]. Briefly, feces were diluted in PBS (0.2 *M* phosphate-buffered saline, pH 7.2), homogenized, filtered through sterile four-ply cotton gauze, pelleted for 5 min at 300 x *g,* resuspended in molecular grade water and layered on top of a 1.5 *M* sucrose solution. After centrifugation for 10 min at 1,700 x *g* the interphase was collected, and the process was repeated with a 0.75 *M* sucrose gradient. The resulting pellet was collected, washed in PBS, and resuspended in 2 ml of molecular-grade water. 0.2 ml of the resulting sample was used as starting material for phenol: chloroform: isoamyl alcohol (25: 24: 1) DNA extraction, eluted in IDTE buffer and stored at -20 °C.

Non-human primate fecal samples in 1:1 RNAlater nucleic acid preservation solution (ThermoFisher) were thawed on ice and homogenized by vortexing prior to transferring 0.2 g of homogenate to bead beating tubes (for a total of 0.1 g fecal material) for extraction using the

Qiagen DNeasy PowerLyzer PowerSoil kit. gDNA was eluted in Qiagen C6 buffer and stored at -20 °C.

Metabarcoding See **Protocol** for step-by-step instructions. For compatibility of sequencing libraries across primer sets and amplicon library types, we created a 2-step Illumina Nextera-based protocol that does not require custom sequencing primers to be added to the sequencing cartridge. Primers for the first (amplicon) PCR consist of a locus-specific sequence (see **Protocol** and **Supp Info Table 1** for locus-specific primer sequences) followed by the Nextera adapter sequences: F-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG and R-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG. A second, limited cycle (indexing) PCR was then used to add Nextera indexing primers to both ends. Note that Platinum II MasterMix (ThermoFisher) has a universal annealing temperature of 60 °C regardless of primer melting temperature. PCRs were run in triplicate with the following conditions: ThermoFisher 1 X Platinum II Hot Start PCR MasterMix, 0.2 µM forward primer with Nextera adapter, 0.2 µM reverse primer with Nextera adapter, 0.2 X ThermoFisher Platinum II GC Enhancer, 0.8 ng/µl gDNA in a total 12.5 µl reaction; 94 °C for 2 minutes, 30 cycles of [94 °C for 15 seconds, 60 °C for 15 seconds, 68 °C for 15seconds], and hold at 4 °C. Triplicate reactions were then pooled and amplicons were cleaned using Ampure XP beads (Beckman Coulter, Brea, California, USA) then used as template for indexing PCR as follows: 1 X KAPA HiFi HotStart ReadyMix (Roche, Basel, Switzerland), 1 X Nextera Unique Dual Index primers (Illumina, San Diego, California, USA), 1 µl of clean amplicons in a total 12.5 µl reaction; 95 °C for 3 minutes, 10 cycles of [95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds], 72 °C for 5 minutes, and hold at 4 °C. Indexed libraries were cleaned using Ampure XP beads (Beckman Coulter) assessed for concentration on a Qubit fluorometer (ThermoFisher), and pooled for sequencing on an Illumina MiSeq with 300 x 300 cycle chemistry using default index and sequencing read primers and 10 – 20 % PhiX.

Data processing and Bioinformatics We processed reads from our final two VESPA data sets with both QIIME 2[83] and DADA2 v.1.16.0[84] in the R environment v.3.6.3 and found that, while results were similar, DADA2 was more user-friendly (*i.e.,* did not require installation of new software, required less steps, and was implementable within a familiar computing environment). Read files were converted to vectors and filtered for quality using the filterAndTrim command with default settings plus modifiers to remove primers (trimLeft = c(18,20)), residual PhiX reads (rm.phix = TRUE), and short sequences (minLen = 100). Error rate for forward and reverse reads were calculated using the learnErrors command, data were dereplicated using the derepFastq command, and Sequence Variants were inferred using the dada command. Read pairs were merged using the mergePairs command with justConcatenate = TRUE and chimeras were removed using the removeBimeraDenovo command with default parameters. Taxonomy assignments were made using the assignTaxonomy command and the PR2 version 4.14.0 database, which contains 18S and 16S sequences at species-level resolution. For comparison we also tested 2 other taxonomy databases: v132 which includes all eukaryotic organisms from the SILVA v132 database and v128 which includes all eukaryotic organisms from the SILVA v128 database plus corrected species labels for *Blastocystis* and additional *Entamoeba* sequences. However, we found that the PR2 database returned higher numbers of fully assigned ASVs. Any ASVs not assigned taxonomy using the PR2 database were queried against the full NCBI nucleotide database on September 3rd, 2022 using MegaBLAST[85] with default parameters.

Data Availability

Sequence data that support the findings of this study have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive with BioProject ID PRJNA944233 and BioSample accessions SAMN33744948 to SAMN33744999. Source data are provided with this paper.

References

1. Whipps JML, K.; Cooke, R.C. Mycoparasitism and plant disease control. In: Burge MN, ed. *Fungi in biological control systems*. Manchester University Press; 1988:161-187:chap 9.

2. Konopka A. What is microbial community ecology? *Isme J*. Nov 2009;3(11):1223-30. doi:10.1038/ismej.2009.88

3. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med*. Apr 10 2018;24(4):392-400. doi:10.1038/nm.4517

4. Pepper JW, Rosenfeld S. The emerging medical ecology of the human gut microbiome. *Trends Ecol Evol*. Jul 2012;27(7):381-4. doi:10.1016/j.tree.2012.03.002

5. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. *Cell*. Mar 16 2012;148(6):1258-70. doi:10.1016/j.cell.2012.01.035

6. Laforest-Lapointe I, Arrieta MC. Microbial eukaryotes: a missing link in gut microbiome studies. *mSystems*. Mar-Apr 2018;3(2)doi:10.1128/mSystems.00201-17

7. Kodio A, Menu E, Ranque S. Eukaryotic and prokaryotic microbiota interactions. *Microorganisms*. Dec 17 2020;8(12)doi:10.3390/microorganisms8122018

8. Avramenko RW, Redman EM, Lewis R, Yazwinski TA, Wasmuth JD, Gilleard JS. Exploring the gastrointestinal "nemabiome": deep amplicon sequencing to quantify the species composition of parasitic nematode communities. *Plos One*. Dec 2 2015;10(12)

9. Matijašić M, Meštrović T, Paljetak HC, Perić M, Barešić A, Verbanac D. Gut microbiota beyond bacteria-mycobiome, virome, archaeome, and eukaryotic parasites in IBD. *Int J Mol Sci*. Apr 11 2020;21(8)doi:10.3390/ijms21082668

10. Kohler JR, Hube B, Puccia R, Casadevall A, Perfect JR. Fungi that infect humans. *Microbiol Spectr*. Jun 2017;5(3)doi:10.1128/microbiolspec.FUNK-0014-2016

11. Tedersoo L, Bahram M, Zinger L, et al. Best practices in metabarcoding of fungi: From experimental design to results. *Mol Ecol*. May 2022;31(10):2769-2795. doi:10.1111/mec.16460

12. Vossbrinck CR, Debrunner-Vossbrinck BA. Molecular phylogeny of the Microsporidia: ecological, ultrastructural and taxonomic considerations. *Folia Parasitol (Praha)*. May 2005;52(1-2):131-42; discussion 130. doi:10.14411/fp.2005.017

13. Dobell C. The discovery of the intestinal protozoa of man. *Proc R Soc Med*. 1920;13(Sect Hist Med):1-15.

14. Momčilović S, Cantacessi C, Arsić-Arsenijević V, Otranto D, Tasić-Otašević S. Rapid diagnosis of parasitic diseases: current scenario and future needs. *Clin Microbiol Infect*. Mar 2019;25(3):290-309. doi:10.1016/j.cmi.2018.04.028

15. Ricciardi A, Ndao M. Diagnosis of parasitic infections: what's going on? *J Biomol Screen*. Jan 2015;20(1):6-21. doi:10.1177/1087057114548065

16. Nadler SA, GP DEL. Integrating molecular and morphological approaches for characterizing parasite cryptic species: implications for parasitology. *Parasitology*. Nov 2011;138(13):1688-709. doi:10.1017/S003118201000168X

17. Jackson TF. *Entamoeba histolytica* and *Entamoeba dispar* are distinct species; clinical, epidemiological and serological evidence. *Int J Parasitol*. Jan 1998;28(1):181-6. doi:10.1016/s0020-7519(97)00177-x

18. Fotedar R, Stark D, Beebe N, Marriott D, Ellis J, Harkness J. PCR detection of *Entamoeba histolytica*, *Entamoeba dispar*, and *Entamoeba moshkovskii* in stool samples from Sydney, Australia. *J Clin Microbiol*. Mar 2007;45(3):1035-7. doi:10.1128/JCM.02144-06

19. Cristescu ME. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends Ecol Evol*. Oct 2014;29(10):566-71. doi:10.1016/j.tree.2014.08.001

20. D'Amore R, Ijaz UZ, Schirmer M, et al. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*. Jan 14 2016;17:55. doi:10.1186/s12864-015-2194-9

21. Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat Rev Microbiol*. Jan 2019;17(2):95-109. doi:10.1038/s41579-018-0116-y

22. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One*. Jul 27 2009;4(7):e6372. doi:10.1371/journal.pone.0006372

23. Parfrey LW, Walters WA, Lauber CL, et al. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Frontiers in Microbiology*. Jun 19 2014;5

24. Mann AE, Mazel F, Lemay MA, et al. Biodiversity of protists and nematodes in the wild nonhuman primate gut. *Isme J*. Feb 2020;14(2):609-622. doi:10.1038/s41396-019-0551-4

25. Maritz JM, Rogers KH, Rock TM, et al. An 18S rRNA workflow for characterizing protists in sewage, with a focus on zoonotic trichomonads. *Microb Ecol*. Nov 2017;74(4):923-936. doi:10.1007/s00248-017-0996-9

26. Jarman SN, McInnes JC, Faux C, et al. Adelie penguin population diet monitoring by analysis of food DNA in scats. *Plos One*. Dec 16 2013;8(12)doi:10.1371/journal.pone.0082227

27. Bhadury P, Austen MC. Barcoding marine nematodes: an improved set of nematode 18S rRNA primers to overcome eukaryotic co-interference. *Hydrobiologia*. Mar 2010;641(1):245-251. doi:10.1007/s10750-009-0088-z

28. Avramenko RW, Bras A, Redman EM, et al. High species diversity of trichostrongyle parasite communities within and between Western Canadian commercial and conservation bison herds revealed by nemabiome metabarcoding. *Parasites & Vectors*. May 15 2018;11doi:10.1186/s13071-018-2880-y

29. Avramenko RW, Redman EM, Lewis R, et al. The use of nemabiome metabarcoding to explore gastro-intestinal nematode species diversity and anthelmintic treatment effectiveness in beef calves. *International Journal for Parasitology*. Nov 2017;47(13):893-902. doi:10.1016/j.ijpara.2017.06.006

30. Poissant J, Gavriliuc S, Bellaw J, et al. A repeatable and quantitative DNA metabarcoding assay to characterize mixed strongyle infections in horses. *Int J Parasitol*. Feb 2021;51(2-3):183-192. doi:10.1016/j.ijpara.2020.09.003

31. Dollive S, Peterfreund GL, Sherrill-Mix S, et al. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol*. Jul 3 2012;13(7):R60. doi:10.1186/gb-2012-13-7-r60

32. Krogsgaard LR, Andersen LO, Johannesen TB, et al. Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome. *Clin Transl Gastroenterol*. Jun 19 2018;9(6):161. doi:10.1038/s41424-018-0027-2

33. Gogarten JF, Calvignac-Spencer S, Nunn CL, et al. Metabarcoding of eukaryotic parasite communities describes diverse parasite assemblages spanning the primate phylogeny. *Mol Ecol Resour*. Jan 2020;20(1):204-215. doi:10.1111/1755-0998.13101

34. Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI. How quantitative is metabarcoding: a meta-analytical approach. *Molecular Ecology*. 2019;28(2):420-430.

35. Sergaki C, Anwar S, Fritzsche M, et al. Developing whole cell standards for the microbiome field. *Microbiome*. Aug 9 2022;10(1):123. doi:10.1186/s40168-022-01313-z

36. Marquina D, Andersson AF, Ronquist F. New mitochondrial primers for metabarcoding of insects, designed and evaluated using *in silico* methods. *Mol Ecol Resour*. Jan 2019;19(1):90-104. doi:10.1111/1755-0998.12942

37. Bradley IM, Pinto AJ, Guest JS. Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Applied and Environmental Microbiology*. Oct 2016;82(19):5878-5891. doi:10.1128/Aem.01630-16

38. Bates ST, Berg-Lyons D, Lauber CL, Walters WA, Knight R, Fierer N. A preliminary survey of lichen associated eukaryotes using pyrosequencing. *Lichenologist*. Jan 2012;44(1):137-146. doi:10.1017/S0024282911000648

39. Bower SM, Carnegie RB, Goh B, Jones SR, Lowe GJ, Mak MW. Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. *J Eukaryot Microbiol*. May-Jun 2004;51(3):325-32. doi:10.1111/j.1550-7408.2004.tb00574.x

40. Cavalier-Smith T, Lewis R, Chao EE, Oates B, Bass D. *Helkesimastix marina* n. sp. (Cercozoa: Sainouroidea superfam. n.) a gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist*. Aug 2009;160(3):452-79. doi:10.1016/j.protis.2009.03.003

41. Bråte J, Klaveness D, Rygh T, Jakobsen KS, Shalchian-Tabrizi K. Telonemia-specific environmental 18S rDNA PCR reveals unknown diversity and multiple marine-freshwater colonizations. *BMC Microbiol*. Jun 9 2010;10:168. doi:10.1186/1471-2180-10-168

42. Geisen S, Snoek LB, ten Hooven FC, et al. Integrating quantitative morphological and qualitative molecular methods to analyse soil nematode community responses to plant range expansion. *Methods in Ecology and Evolution*. 2018;9(6):1366-1378.

43. Comeau AM, Li WK, Tremblay JE, Carmack EC, Lovejoy C. Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS One*. 2011;6(11):e27492. doi:10.1371/journal.pone.0027492

44. DeMone C, Hwang M-H, Feng Z, et al. Application of next generation sequencing for detection of protozoan pathogens in shellfish. *Food and waterborne parasitology*. 2020;21:e00096.

45. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *Plos One*. Feb 7 2014;9(2)doi: 10.1371/journal.pone.0087624

46. Hugerth LW, Muller EE, Hu YO, et al. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One*. 2014;9(4):e95567. doi:10.1371/journal.pone.0095567

47. Machida RJ, Knowlton N. PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS One*. 2012;7(9):e46180. doi:10.1371/journal.pone.0046180

48. Sikder M, Vestergård M, Sapkota R, Kyndt T, Nicolaisen M. Evaluation of Metabarcoding Primers for Analysis of Soil Nematode Communities. *Diversity*. 2020;12(10):388.

49. Stoeck T, Behnke A, Christen R, et al. Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol*. Nov 3 2009;7:72. doi:10.1186/1741-7007-7-72

50. Wood JR. DNA barcoding of ancient parasites. *Parasitology*. Apr 2018;145(5):646-655. doi:10.1017/S0031182018000380

51. Zhan AB, Hulak M, Sylvester F, et al. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods in Ecology and Evolution*. Jun 2013;4(6):558-565. doi:10.1111/2041-210x.12037

52. Beermann AJ, Werner MT, Elbrecht V, Zizka VMA, Leese F. DNA metabarcoding improves the detection of multiple stressor responses of stream invertebrates to increased salinity, fine sediment deposition and reduced flow velocity. *Sci Total Environ*. Jan 1 2021;750:141969. doi:10.1016/j.scitotenv.2020.141969

53. Bohmann K, Elbrecht V, Caroe C, et al. Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Mol Ecol Resour*. May 2022;22(4):1231-1246. doi:10.1111/1755-0998.13512

54. Song F, Kuehl JV, Chandran A, Arkin AP. A simple, cost-effective, and automation-friendly direct PCR approach for bacterial community analysis. *mSystems*. Oct 26 2021;6(5):e0022421. doi:10.1128/mSystems.00224-21

55. Albaina A, Aguirre M, Abad D, Santos M, Estonba A. 18S rRNA V9 metabarcoding for diet characterization: a critical evaluation with two sympatric zooplanktivorous fish species. *Ecol Evol*. Mar 2016;6(6):1809-24. doi:10.1002/ece3.1986

56. Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci Rep*. Dec 15 2017;7(1):17668. doi:10.1038/s41598-017-17333-x

57. Deagle BE, Thomas AC, McInnes JC, et al. Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Mol Ecol*. Jan 2019;28(2):391-406. doi:10.1111/mec.14734

58. Yu Z, Ito SI, Wong MK, et al. Comparison of species-specific qPCR and metabarcoding methods to detect small pelagic fish distribution from open ocean environmental DNA. *PLoS One*. 2022;17(9):e0273670. doi:10.1371/journal.pone.0273670

59. Weerakoon KG, McManus DP. Cell-free DNA as a diagnostic tool for human parasitic infections. *Trends Parasitol*. May 2016;32(5):378-391. doi:10.1016/j.pt.2016.01.006

60. Maldonado A, Simões RO, Luiz JS, Costa-Neto SF, Vilela RV. A new species of *Physaloptera* (Nematoda: Spirurida) from *Proechimys gardneri* (Rodentia: Echimyidae) from the Amazon rainforest and molecular phylogenetic analyses of the genus. *J Helminthol*. Jul 24 2019;94:e68. doi:10.1017/S0022149X19000610

61. Abraham JS, Sripoorna S, Maurya S, Makhija S, Gupta R, Toteja R. Techniques and tools for species identification in ciliates: a review. *Int J Syst Evol Microbiol*. Apr 2019;69(4):877-894. doi:10.1099/ijsem.0.003176

62. Macheriotou L, Guilini K, Bezerra TN, et al. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecol Evol*. Feb 2019;9(3):1211-1226. doi:10.1002/ece3.4814

63. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. Jan 2013;41(Database issue):D590-6. doi:10.1093/nar/gks1219

64. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*. Jan 4 2018;46(D1):D41-D47. doi:10.1093/nar/gkx1094

65. Yilmaz P, Parfrey LW, Yarza P, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*. Jan 2014;42(Database issue):D643-8. doi:10.1093/nar/gkt1209

66. Centers for Disease Control GH, Division of Parasitic Diseases and Malaria. Alphabetical Index of Parasitic Diseases. https://www.cdc.gov/parasites/az/index.html

67. Lukeš J, Stensvold CR, Jirků-Pomajbiková K, Parfrey LW. Are human intestinal eukaryotes beneficial or commensals? *Plos Pathogens*. Aug 2015;11(8). doi:0.1371/journal.ppat.1005039

68. Modrý DP, B. Petrželková, K. Hasegawa, H. *Parasites of apes an atlas of coproscopic diagnostics*. vol 78. Frankfurt Contributions to Natural History / Frankfurter Beiträge zur Naturkunde. Edition Chimaira; 2018:198.

69. Taylor MA, Coop RL, Wall R. *Veterinary parasitology*. 4th edition. ed. John Wiley and Sons, Inc.; 2016:p.

70. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-7. doi:10.1093/nar/gkh340

71. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Mol Biol Evol*. Jun 25 2021;38(7):3022-3027. doi:10.1093/molbev/msab120

72. Ludwig W, Strunk O, Westram R, et al. ARB: a software environment for sequence data. *Nucleic Acids Res*. 2004;32(4):1363-71. doi:10.1093/nar/gkh293

73. Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res*. Nov 2011;39(21):e145. doi:10.1093/nar/gkr732

74. Klindworth A, Pruesse E, Schweer T, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. Jan 7 2013;41(1):e1. doi:10.1093/nar/gks808

75. Loakes D. Survey and summary: The applications of universal DNA base analogues. *Nucleic Acids Res*. Jun 15 2001;29(12):2437-47. doi:10.1093/nar/29.12.2437

76. Levin JD, Fiala D, Samala MF, Kahn JD, Peterson RJ. Position-dependent effects of locked nucleic acid (LNA) on DNA sequencing and PCR primers. *Nucleic Acids Res*. 2006;34(20):e142. doi:10.1093/nar/gkl756

77. Wang C, Zhang T, Wang Y, Katz LA, Gao F, Song W. Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. *Proc Biol Sci*. Jul 26 2017;284(1859) doi:10.1098/rspb.2017.0425

78. Piñol J, Senar MA, Symondson WO. The choice of universal primers and the chatacteristics of the species mixtuer determine when DNA metabarcoding can be quantitative. *Molecular Ecology Notes*. 2018;28:407-419.

79. Clemente JC, Pehrsson EC, Blaser MJ, et al. The microbiome of uncontacted Amerindians. *Sci Adv*. Apr 3 2015;1(3)doi:10.1126/sciadv.1500183

80. Friant S, Ziegler TE, Goldberg TL. Changes in physiological stress and behaviour in semi-free-ranging red-capped mangabeys (*Cercocebus torquatus*) following antiparasitic treatment. *Proceedings of the Royal Society B-Biological Sciences*. Jul 27 2016;283(1835)

81. Friant S, Ziegler TE, Goldberg TL. Primate reinfection with gastrointestinal parasites: behavioural and physiological predictors of parasite acquisition. *Anim Behav*. Jul 2016;117:105-113. doi:10.1016/j.anbehav.2016.04.006

82. Walderich B, Müller L, Bracha R, Knobloch J, Burchard GD. A new method for isolation and differentiation of native *Entamoeba histolytica* and *E. dispar cysts* from fecal samples. *Parasitol Res*. 1997;83(7):719-21. doi:10.1007/s004360050326

83. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. May 2010;7(5):335-6. doi:10.1038/nmeth.f.303

84. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. Jul 2016;13(7):581-3. doi:10.1038/nmeth.3869

85. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. Database indexing for production MegaBLAST searches. *Bioinformatics*. Aug 15 2008;24(16):1757-64. doi:10.1093/bioinformatics/btn322

Acknowledgements

We gratefully acknowledge Lyric Bartholomay, Timothy Yoshino, Peter Thompson, Monica Santin-Duran, Marie Pinkerton, Cecilia Escobar Lopez, Mostafa Zamanian, and Johanna Elfenbein for assistance in obtaining specimens, for cloning reagents, and for technical assistance. We thank Laura Parfrey for kindly providing the v132 and v128 taxonomy databases. This research was funded by National Institutes of Health (NIH) via grants 1R01AG049395-01 and 1R21AI163592-01 and the UW Madison Parasitology and Vector Biology Training Grant T32AI007414.

Author Contributions

T.L.G., S.R.F., L.A.O., conceived the study. T.L.G. and L.A.O. designed the study. L.A.O. collected and analyzed data. L.A.O. and T.L.G. wrote the manuscript. S.R.F. and B.M.G. provided data. S.R.F., B.M.G., L.K.K., M.C., O.N., and M.D.-B. provided samples. All authors made substantive intellectual contributions, revised the manuscript, and approved the final draft.

Figures

Figure 1.

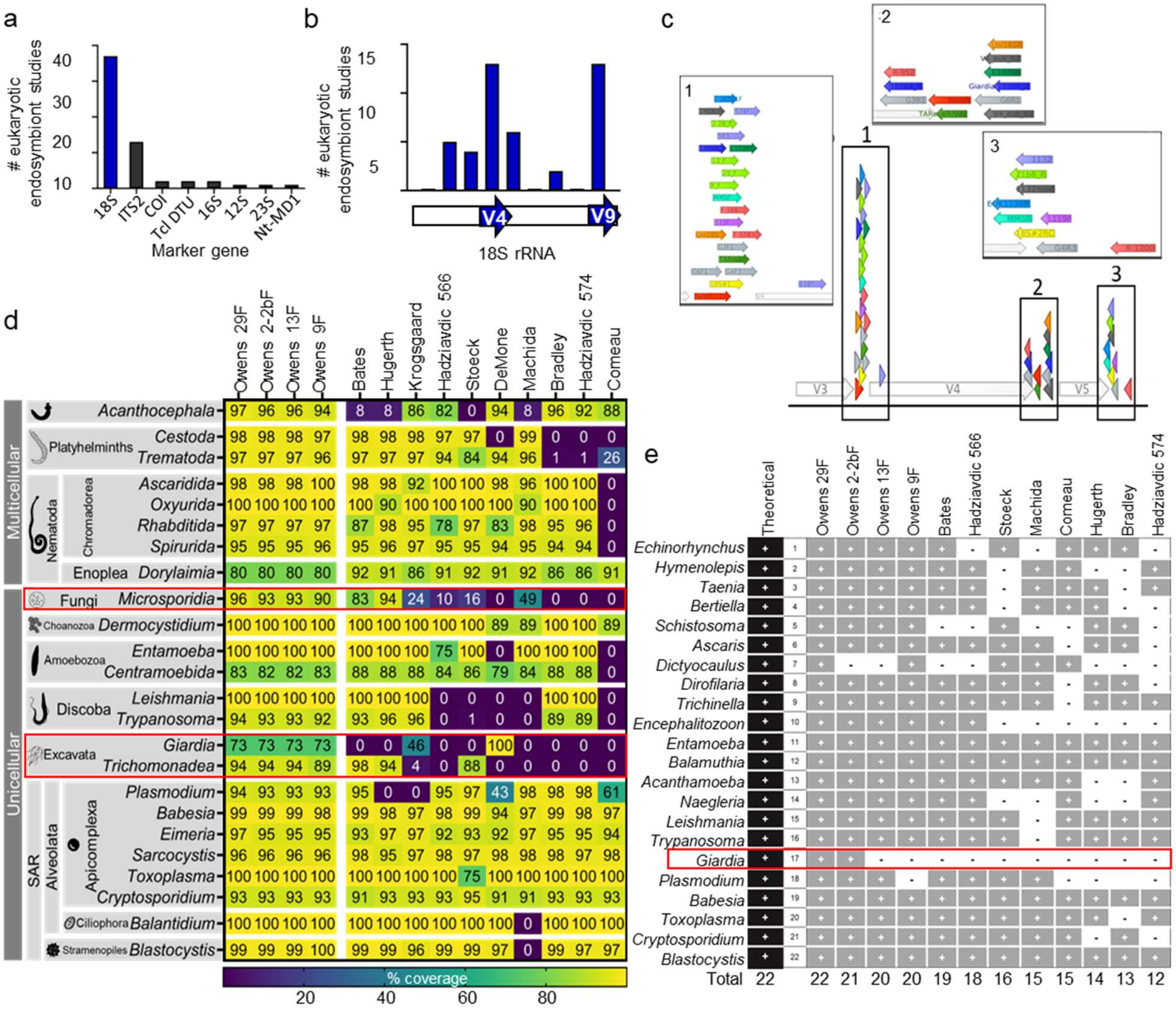
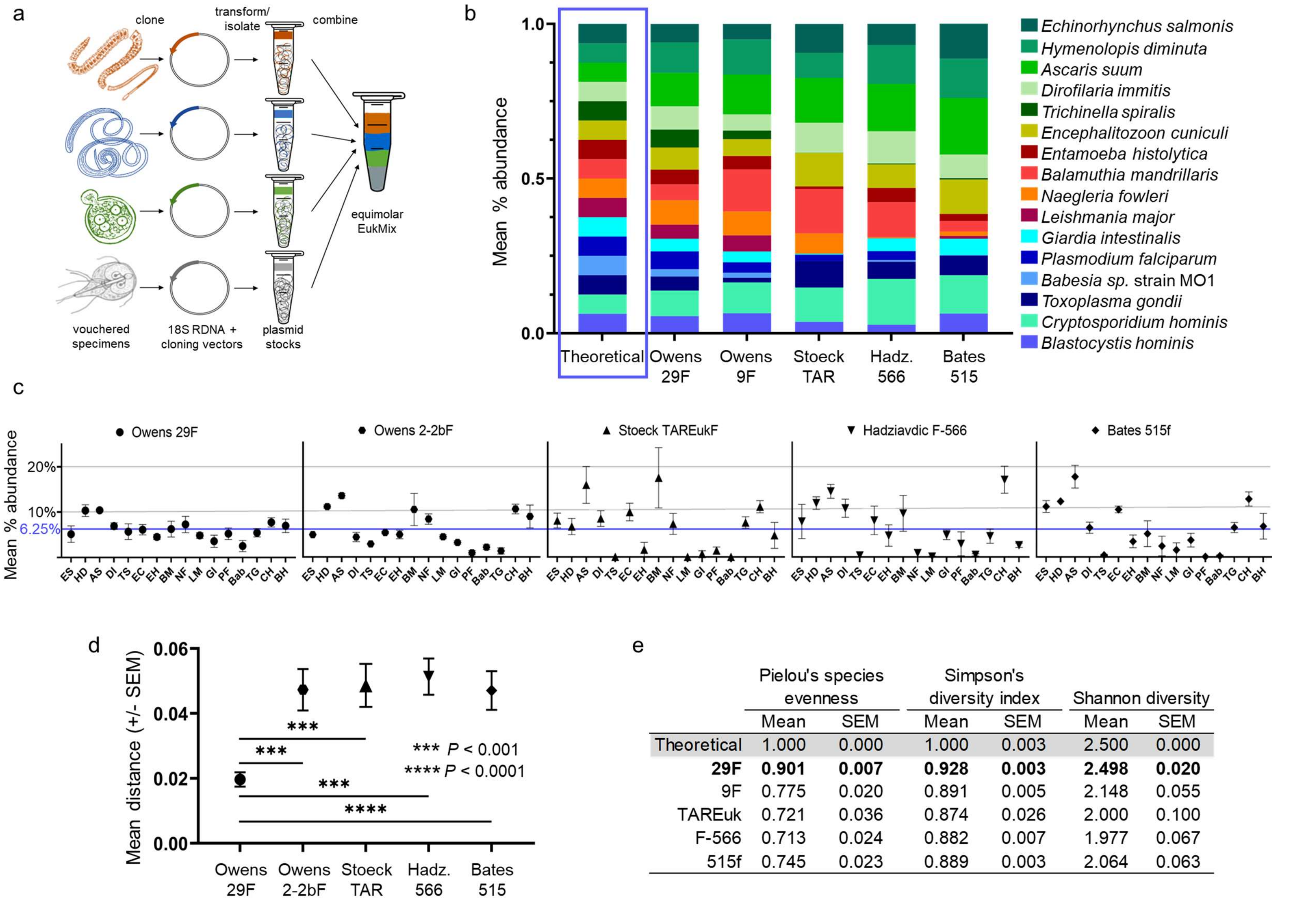


Figure 1. VESPA development and evaluation. a, Histogram of marker genes identified in a literature review of 54 host-associated eukaryotic endosymbiont studies (see **Supp File**). **b,** 18S rRNA primer sets from our literature review shown as a histogram binned by location along the 18S gene. Hypervariable regions V4 and V9 are demarcated by blue arrows below the x-axis. **c,** Generalized map of hypervariable regions V3 – V5 (open arrows) of eukaryotic 18S SSU rRNA gene. Newly designed and published metabarcoding primer sets are shown as colored arrows and boxed areas 1 -3 are expanded as insets. See **Supp Info Table 1** for full primer names and sequences. **d,** Heat map of published and new 18S V4 primer set coverage across clades exclusively containing parasites of vertebrates. Percent overall complementarity (% coverage) is shown as numbers and as a color scale (color key below heatmap) with taxonomic labels to the left. Red boxes highlight clades with low overall ("problematic") coverage. **e,** Vertebrate endoparasite PCR panel showing amplification (+) or lack of amplification (-) of single-organism gDNA templates across new and published primer sets. Total represents the number of successful amplifications per primer out of 22 possible, shown in left-most "Theoretical" column. Red box highlights clade with low overall ("problematic") amplification.

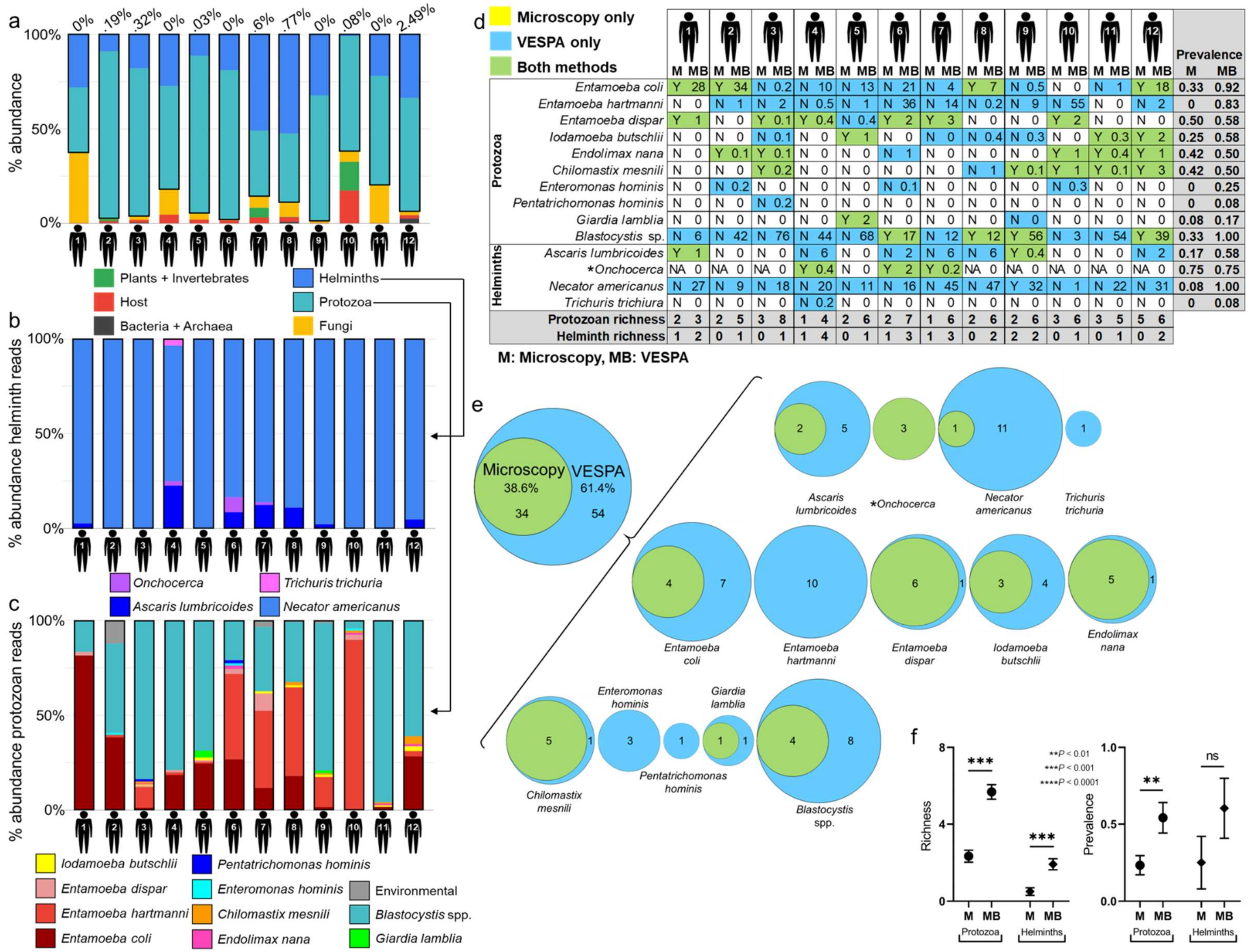
Figure 2.



	Pielou's species evenness		Simpson's diversity index		Shannon diversity	
	Mean	SEM	Mean	SEM	Mean	SEM
Theoretical	1.000	0.000	1.000	0.003	2.500	0.000
29F	**0.901**	**0.007**	**0.928**	**0.003**	**2.498**	**0.020**
9F	0.775	0.020	0.891	0.005	2.148	0.055
TAREuk	0.721	0.036	0.874	0.026	2.000	0.100
F-566	0.713	0.024	0.882	0.007	1.977	0.067
515f	0.745	0.023	0.889	0.003	2.064	0.063

Figure 2. Testing metabarcoding methods for amplification bias using a community standard. a, Schematic overview of EukMix creation via 18S isolation and cloning. **b,** Equimolar EukMix community standard metabarcoding across primer sets as compared to theoretical input (leftmost bar, blue box) shown as % abundance of reads per organism. **c,** Equimolar EukMix community standard metabarcoding. Reads assigned to each component organism are shown as mean % abundance of three replicates +/- standard error of the mean (SEM) with theoretical input level of 6.25 % displayed as a blue horizontal line. ES, *Echinorhynchus salmonis;* HD, *Hymenolepis diminuta;* AS, *Ascaris suum;* DI, *Dirofilaria immitis;* TS, *Trichinella spiralis;* EC, *Encephalitozoon cuniculi;* EH, *Entamoeba histolytica;* BM, *Balamuthia mandrillaris,* NF, *Naegleria fowleri;* LM, *Leishmania major;* GI, *Giardia intestinalis,* PF, *Plasmodium falciparum;* Bab, *Babesia* sp. strain MO1; TG, *Toxoplasma gondii;* CH, *Cryptosporidium hominis;* BH, *Blastocystis hominis.* See **Supp Info Table 2** for parasite sources and strains. **d,** Mean absolute distance to the theoretical input level for each primer set for three replicates +/- SEM. *P* values are derived from two-tailed Wilcoxon matched-pairs signed rank tests. Owens 29F was significantly different from all other primer sets (shown as bars with asterisks). All comparisons not shown are not significant. **e,** Diversity metrics based on EukMix analysis compared to theoretically equal input (shaded row). Primer set 29F represented the underlying community most accurately by all three metrics (bolded row).

Figure 3.



d

🟨 Microscopy only

🟦 VESPA only

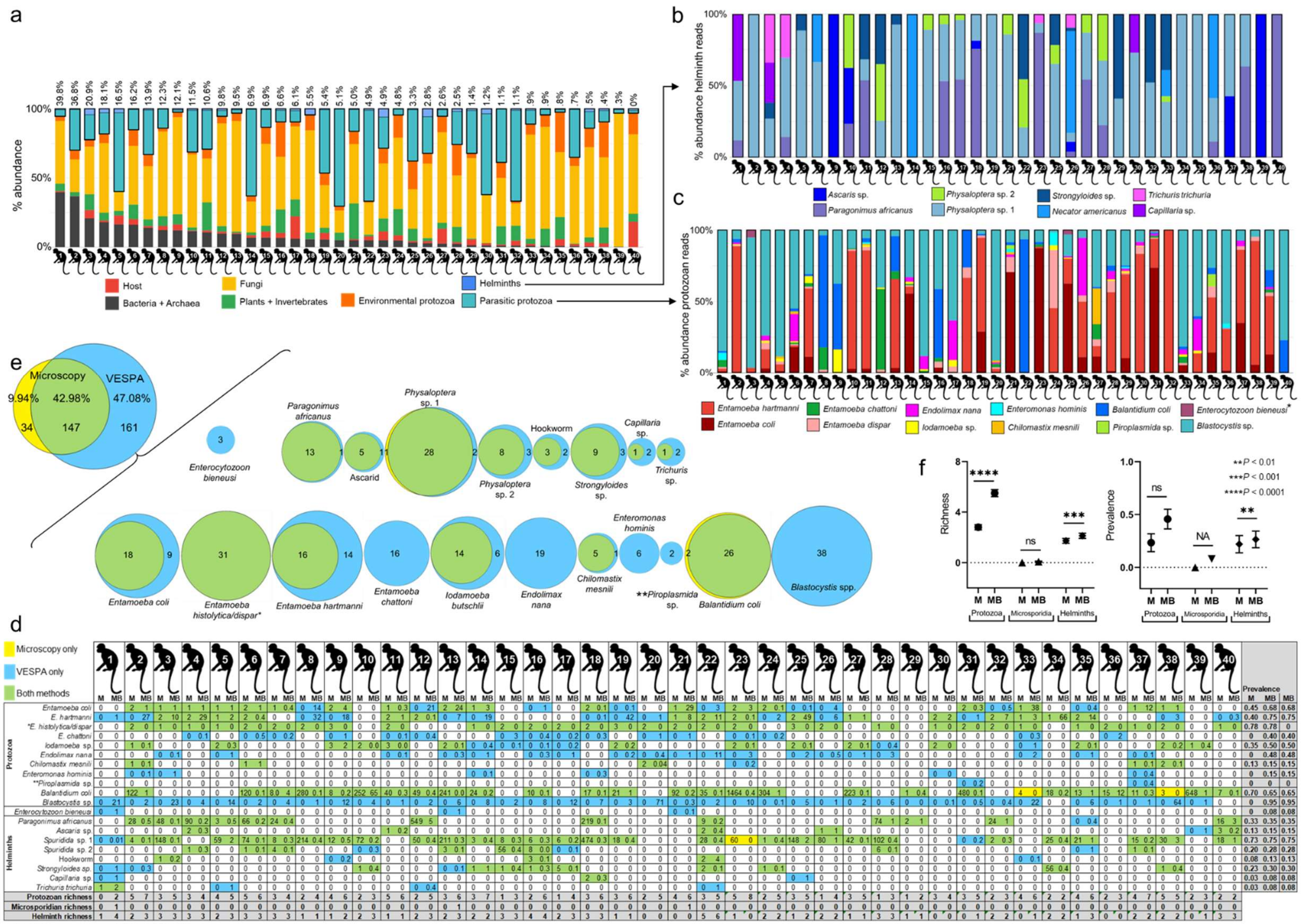
🟩 Both methods

		1		2		3		4		5		6		7		8		9		10		11		12		Prevalence	
		M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB	M	MB
Protozoa	*Entamoeba coli*	Y	28	Y	34	N	0.2	N	10	N	13	N	21	N	4	Y	7	N	0.5	N	0	N	1	Y	18	0.33	0.92
	Entamoeba hartmanni	N	0	N	1	N	2	N	0.5	N	1	N	36	N	14	N	0.2	N	9	N	55	N	0	N	2	0	0.83
	Entamoeba dispar	Y	1	N	0	Y	0.1	Y	0.4	N	0.4	Y	2	Y	3	N	0	N	0	Y	2	N	0	N	0	0.50	0.58
	Iodamoeba butschlii	N	0	N	0	N	0.1	N	0	Y	1	N	0	N	0	N	0.4	N	0.3	N	0	Y	0.3	Y	2	0.25	0.58
	Endolimax nana	N	0	Y	0.1	Y	0.1	N	0	N	0	N	1	N	0	N	0	N	0	Y	1	Y	0.4	Y	1	0.42	0.50
	Chilomastix mesnili	N	0	N	0	Y	0.2	N	0	N	0	N	0	N	0	N	1	Y	0.1	Y	1	Y	0.1	Y	3	0.42	0.50
	Enteromonas hominis	N	0	N	0.2	N	0	N	0	N	0	N	0.1	N	0	N	0	N	0	N	0.3	N	0	N	0	0	0.25
	Pentatrichomonas hominis	N	0	N	0	N	0.2	N	0	N	0	N	0	N	0	N	0	N	0	N	0	N	0	N	0	0	0.08
	Giardia lamblia	N	0	N	0	N	0	N	0	Y	2	N	0	N	0	N	0	N	0	N	0	N	0	N	0	0.08	0.17
	Blastocystis sp.	N	6	N	42	N	76	N	44	N	68	Y	17	N	12	Y	12	Y	56	N	3	N	54	Y	39	0.33	1.00
Helminths	*Ascaris lumbricoides*	Y	1	N	0	N	0	N	6	N	0	N	2	N	6	N	6	Y	0.4	N	0	N	0	N	2	0.17	0.58
	**Onchocerca*	NA	0	NA	0	NA	0	Y	0.4	N	0	Y	2	Y	0.2	NA	0	NA	0	NA	0	NA	0	NA	0	0.75	0.75
	Necator americanus	N	27	N	9	N	18	N	20	N	11	N	16	N	45	N	47	Y	32	N	1	N	22	N	31	0.08	1.00
	Trichuris trichiura	N	0	N	0	N	0	N	0.2	N	0	N	0	N	0	N	0	N	0	N	0	N	0	N	0	0	0.08
	Protozoan richness	**2**	**3**	**2**	**5**	**3**	**8**	**1**	**4**	**2**	**6**	**2**	**7**	**1**	**6**	**2**	**6**	**2**	**6**	**3**	**6**	**3**	**5**	**5**	**6**		
	Helminth richness	**1**	**2**	**0**	**1**	**0**	**1**	**1**	**4**	**0**	**1**	**1**	**3**	**1**	**3**	**0**	**2**	**2**	**2**	**0**	**1**	**0**	**1**	**0**	**2**		

M: Microscopy, MB: VESPA

Figure 3. VESPA compared to microscopy in human clinical samples. a, **b**, **c**, VESPA metabarcoding data. VESPA data are shown as percent relative abundance of each organism category with (**a**) all quality-filtered reads included, (**b**) with helminth reads only, or (**c**) with protozoal reads only (archaea, bacteria, host, plants, invertebrates, and fungi removed in **b** and **c**). **a,** Numbers above bars are the total percentage of prokaryotic (bacterial + archaeal) reads. **d,** Microscopy versus VESPA. Microscopy findings (M) are shown as a presence/absence (Y = present, N = absent, NA = not assessed) and VESPA metabarcoding (MB) findings are shown as % abundance of quality-filtered reads. Blue cells represent detection by VESPA, green cells by both VESPA and microscopy, and white cells by neither method. No organisms were identified by microscopy alone. Richness (final 2 rows, shaded cells) is defined as the total number of species detected by the specified method. Prevalence (final 2 columns, shaded cells) is defined as the proportion of the population positive for an organism by the specified method. Note that *Onchocerca* is not detectable in fecal samples by microscopy (asterisk). **e**, Proportional Venn diagrams of findings by microscopy versus VESPA. Individuals identified as positive for the listed organisms by VESPA (blue) or both (green) are shown as numbers in each circle. Overall findings summed over all organisms are shown to the left of the bracket (not to scale). Note that *Onchocerca* is not detectable in fecal samples by microscopy (asterisk). **f,** Richness and prevalence calculations for microscopy (M) and VESPA metabarcoding (MB) findings. Data are shown as mean +/- SEM. *P* values are derived from Wilcoxon matched-pairs signed rank tests, 2-tailed. ns, not significant.

Figure 4.



M: Microscopy, MB: VESPA

Figure 4. VESPA compared to microscopy in non-human primate clinical samples. a, **b**, **c**, VESPA metabarcoding data. VESPA data are shown as percent relative abundance of each organism category with (**a**) all quality-filtered reads included, (**b**) with helminth reads only, or (**c**) with protozoal reads only (archaea, bacteria, host, plants, invertebrates, and fungi removed in **b** and **c**). **a**, Numbers above bars are the total percentage of prokaryotic (bacterial + archaeal) reads. **c**, Asterisk indicates a microsporidian parasite. **d**, Microscopy versus VESPA. Microscopy findings (M) are shown as a qualitative score (1 least – 3 most) for protozoa, larvae/gram feces for *Strongyloides*, and eggs/gram feces for all other helminths. VESPA findings (MB) are shown as % abundance of quality-filtered reads. Yellow cells represent parasite detection by microscopy, blue cells by VESPA, green cells by both methods, and white cells by neither method. Richness (final 2 rows, shaded cells) is defined as the total number of species detected by the specified method. Prevalence (final 2 columns, shaded cells) is defined as the proportion of the population positive for an organism by the specified method. Note that *Entamoeba histolytica* and *Entamoeba dispar* are a cryptic species complex that cannot be resolved by microscopy (asterisk) and *Piroplasmida* sp. are not detectable in fecal samples by microscopy (double asterisk). **e,** Proportional Venn diagrams of findings by microscopy versus VESPA. Individuals identified as positive for the listed organisms by microscopy (yellow), VESPA (blue), or both (green) are shown as numbers in each circle. Overall findings summed over all organisms are shown to the left of the bracket (not to scale). Note that *Entamoeba histolytica* and *Entamoeba dispar* are a cryptic species complex that cannot be resolved by microscopy (asterisk) and *Piroplasmida* sp. are not detectable in fecal samples by microscopy (double asterisk). **f,** Richness and prevalence calculations for microscopy (M) and VESPA (MB)

findings. Data are shown as mean +/- SEM. *P* values are derived from Wilcoxon matched-pairs signed rank tests, 2-tailed. ns, not significant. NA, not applicable (single data point only).

Supplementary Information

Supplementary Table 1. 18S primers used in this study.

Primer Name	Reference	F/R	Sequence (5′ – 3′)	Region
515f	Bates 2012	F	GTGCCAGCMGCCGCGGTAA	V4
1119r	Bates 2012	R	GGTGCCCTTCCGTCA	V4
18S-EUK581-F	Bower 2004	F	GTGCCAGCAGCCGCG	V4
18S-EUK1134-R	Bower 2004	R	TTTAAGTTTCAGCCTTGCG	V4
TAReuk454FWD1	Stoeck 2010	F	CCAGCASCYGCGGTAATTCC	V4
V4r	Bradley 2016	R	ACTTTCGTTCTTGAT	V4
3NDf	Cavalier-Smith 2009	F	GGCAAGTCTGGTGCCAG	V4
V4_euk_R2	Brate 2010	R	ACGGTATCTRATCRTCTTCG	V4
V4_euk_R1	Brate 2010	R	GACTACGACGGTATCTRATCRTCTTCG	V4
1132mod	Giesen 2018	R	TCCGTCAATTYCTTTAAGT	V4
E572F	Comeau 2011	F	CYGCGGTAATTCCAGCTC	V4
E1009R	Comeau 2011	R	AYGGTATCTRATCRTCTTYG	V4
18SV4_F	DeMone 2020	F	GCCGCGGTAATTCCAGCTC	V4
18SV4_R	DeMone 2020	R	ATYYTTGGCAAATGCTTTCGC	V4
Giardia 18SV4_R	DeMone 2020	R	ATACGGTGGTGTCTGATCGC	V4
F-566	Hadziavdic 2014	F	CAGCAGCCGCGGTAATTCC	V4
R-1200	Hadziavdic 2014	R	CCCGTGTTGAGTCAAATTAAGC	V4
F-574	Hadziavdic 2014	F	GCGGTAATTCCAGCTCCAA	V4
R-952	Hadziavdic 2014	R	TTGGCAAATGCTTTCGC	V4
574	Hugerth 2014	F	CGGTAAYTCCAGCTCYV	V4

1132	Hugerth 2014	R	CCGTCAATTHCTTYAART	V4
616	Hugerth 2014	F	TTAAARVGYTCGTAGTYG	V4
563	Hugerth 2014	F	GCCAGCAVCYGCGGTAAY	V4
G3F1	Krogsgaard 2018	F	GCCAGCAGCCGCGGTAATTC	V4
G3R1	Krogsgaard 2018	R	ACATTCTTGGCAAATGCTTTCGCAG	V4
G4F3	Krogsgaard 2018	F	AGCCGCGGTAATTCCAGCTC	V4
G4R3	Krogsgaard 2018	R	GGTGGTGCCCTTCCGTCAAT	V4
G6F1	Krogsgaard 2018	F	TGGAGGGCAAGTCTGGTGCC	V4
G6R1	Krogsgaard 2018	R	TACGGTATCTGATCGTCTTCGATCCC	V4
18S#1	Machida 2012	F	CTGGTGCCAGCAGCCGCGGYAA	V4
18S#2RC	Machida 2012	R	TCCGTCAATTYCTTTAAGTT	V4
MMSF	Sikder 2020	F	GGTGCCAGCAGCCGCGGTA	V4
MMSR	Sikder 2020	R	CTTTAAGTTTCAGCTTTGC	V4
Nem18SlongF	Wood 2013	F	CAGGGCAAGTCTGGTGCCAGCAGC	V4
Nem18SlongR	Wood 2013	R	GACTTTCGTTCTTGATTAATGAA	V4
Uni18S	Zhan 2013	F	AGGGCAAKYCTGGTGCCAGC	V4
Uni18SR	Zhan 2013	R	GRCGGTATCTRATCGYCTT	V4
9F	This study	F	CTGGTGCCAGCAGCCGCGG	V4
13F	This study	F	TGGTGCCAGCAGCCGCGG	V4
29F	This study	F	AGCAGCCGCGGTAATTCC	V4
2-2bF	This study	F	TGGTGCCAGCASCCGCG	V4
21b8R	This study	R	TCAATTYCTTIAASTTTC	V4
EukA_F	Medlin 1988	F	AACCTGGTTGATCCTGCCAGT	5' terminus
EukB_R	Medlin 1988	R	TGATCCTTCTGCAGGTTCACCTAC	3' terminus
1520_R	Lopez-Garcia 2003	R	CYGCAGGTTCACCTAC	3' terminus

V3Mod_F	This study (modified from Flaherty 2018)	F	CCGGAGAGRGAGCMTKAG	5' terminus
EukBshort_R	This study (modified from Medlin 1988)	R	CCTTCCGCAGGTTCACCTAC	3' terminus
LAOEukF	This study	F	CTGGTTGATCCTGCCAGTAKT	5' terminus
LAOEuk2F	This study	F	CTGGTTGATCCTGCCAGT	5' terminus
LAO18SF	This study	F	CGCGAANGGCTCATTANAWCAGC	5' terminus
LAOGiarF	This study	F	ACGGCTCAGGACAACGGTT	5' terminus
LAO1498R	This study	R	GGTTCACCTACGGANACCTTGTTA	3' terminus
LAOECR	This study	R	TCGTCTTCTCAGCGCCGGT	3' terminus
LAOEntCrypF	This study	F	GATTAAGCCATGCATGTSTAAG	5' terminus
LAO380F	This study	F	GGTTCGACTCCGGAGAG	5' terminus
LAOTW2F	This study	F	TGGATAACTGTAATRACTCT	5' terminus
LAOTW3R	This study	R	GACCTYACTAAACCATTCAATC	3' terminus

F, Forward primer; R, Reverse primer.

Supplementary Table 2. Parasite specimens and sources.

Organism	Sample type	Source	Catalog #
Echinorhynchus salmonis	Whole adult worms	UW Madison School of Veterinary Medicine, Dr. Tony Goldberg	N/A
Hymenolepis diminuta	Whole adult worms	UW Madison School of Veterinary Medicine, Dr. Timothy Yoshino	N/A
Taenia hydatigena	Cysts	Wisconsin Veterinary Diagnostic Lab	N/A
Bertiella studeri	Proglottids	UW Madison School of Veterinary Medicine, Dr. Tony Goldberg	N/A
Schistosoma mansoni Strain NMRI	DNA	BEI Resources	NR-28911
Ascaris suum	Whole adult worms	Wisconsin Veterinary Diagnostic Lab	N/A
Dictyocaulus viviparous	Whole adult worms	Wisconsin Veterinary Diagnostic Lab	N/A
Dirofilaria immitis Strain Missouri 2005	DNA	BEI Resources	NR-44348
Trichinella spiralis	DNA	USDA Animal Parasitic Diseases Laboratory	N/A
Encephalitozoon cuniculi Strain CDC: V282	DNA	BEI Resources	NR-13510
Entamoeba histolytica Strain HK-9	DNA	BEI Resources	NR-175
Balamuthia mandrillaris CDC: V188	Axenic culture	BEI Resources	NR-46452
Acanthamoeba sp. Strain CDC: 12741:1	DNA	BEI Resources	NR-45611
Naegleria fowleri Strain CDC: V414	Axenic culture	BEI Resources	NR-46494
Leishmania major Strain NIH SD	DNA	BEI Resources	NR-48764

Trypanosoma cruzi Strain G	DNA	BEI Resources	NR-50238
Giardia lamblia Strain WB clone C6	DNA	BEI Resources	NR-15894
Plasmodium falciparum Strain D6	DNA	BEI Resources	MRA-398
Babesia sp. Strain MO1	DNA	BEI Resources	NR-50663
Toxoplasma gondii	DNA	UW Madison Department of Medical Microbiology and Immunology, Dr. Laura Knoll	NR-33509
Cryptosporidium hominis Strain TU502	DNA	BEI Resources	NR-2520
Blastocystis hominis Strain BT1	DNA	ATCC (American Type Culture Collection)	50608

N/A, not applicable.

Supplementary Table 3. Equimolar EukMix components and full-length 18S cloning primers.

	Organism	FWD primer*	REV primer*
1	*Echinorhynchus salmonis*	EukA_F	EukB_R
2	*Hymenolepis diminuta*	LAOTW2F	LAOTW3R
3	*Ascaris suum*	LAO18SF	LAO1498R
4	*Dirofilaria immitis*	LAO18SF	LAO1498R
5	*Trichinella spiralis*	V3mod_F	EukBshort_R
6	*Encephalitozoon cuniculi*	V3mod_F	LAOECR
7	*Entamoeba histolytica*	LAOEuk2F	EukB_R
8	*Balamuthia mandrillaris*	EukA_F	EukB_R
9	*Naegleria fowleri*	LAO380F	LAO1498R
10	*Giardia intestinalis*	LAO380F	EukB_R
11	*Leishmania major*	LAOEukF	EukB_R
12	*Plasmodium falciparum*	EukA_F	EukB_R
13	*Babesia* sp. strain MO1	EukA_F	EukB_R
14	*Toxoplasma gondii*	EukA_F	EukB_R
15	*Cryptosporidium hominis*	EukA_F	LAO1498R
16	*Blastocystis hominis*	LAOEukF	LAO1498R

*See Supplementary table 1 for primer sequences and references.

Supplementary Table 4. Equimolar EukMix metabarcoding accuracy metrics.

		Mean distance from the theoretical by primer set					Estimated abundance pattern
		Owens 29F	Owens 2-2bF	Stoeck TAREuk	Hadziavdic F-566	Bates 515f	
1	*Echinorhynchus salmonis*	-1.12%	-1.36%	1.71%	1.48%	4.77%	mixed
2	*Hymenolepis diminuta*	4.09%	4.75%	0.39%	5.51%	5.88%	over
3	*Ascaris suum*	2.48%	7.13%	9.48%	8.08%	11.28%	over
4	*Dirofilaria immitis*	0.66%	-1.91%	2.14%	4.39%	0.15%	mixed
5	*Trichinella spiralis*	-0.57%	-3.40%	-6.22%	-5.91%	-5.84%	under
6	*Encephalitozoon cuniculi*	-0.14%	-0.92%	3.56%	1.72%	4.12%	mixed
7	*Entamoeba histolytica*	-1.75%	-1.34%	-4.67%	-1.56%	-2.86%	under
8	*Balamuthia mandrillaris*	0.00%	4.11%	11.05%	3.26%	-1.18%	mixed
9	*Naegleria fowleri*	1.03%	2.03%	0.98%	-5.41%	-3.86%	mixed
10	*Leishmania major*	-1.41%	-1.84%	-6.25%	-6.15%	-4.72%	under
11	*Giardia intestinalis*	-2.69%	-3.09%	-5.58%	-1.43%	-2.57%	under
12	*Plasmodium falciparum*	-1.02%	-5.32%	-4.84%	-3.39%	-6.18%	under
13	*Babesia* sp. strain MO1	-0.98%	-4.12%	-6.22%	-5.78%	-6.04%	under
14	*Toxoplasma gondii*	-0.80%	-4.93%	1.27%	-1.73%	0.18%	mixed
15	*Cryptosporidium hominis*	1.48%	4.23%	4.72%	10.62%	6.40%	over
16	*Blastocystis hominis*	0.73%	2.59%	-1.53%	-3.71%	0.45%	mixed

Supplementary Table 5. VESPA MiSeq run metrics.

Library ID	SRA accession	Sample type	Raw reads	Reads post-quality filter	% lost in filter
Human01	SAMN33744948	Human fecal	62,512	56,602	9.45%
Human02	SAMN33744949	Human fecal	32,755	30,195	7.82%
Human03	SAMN33744950	Human fecal	223,911	206,999	7.55%
Human04	SAMN33744951	Human fecal	43,371	39,228	9.55%
Human05	SAMN33744952	Human fecal	116,016	106,130	8.52%
Human06	SAMN33744953	Human fecal	24,095	22,204	7.85%
Human07	SAMN33744954	Human fecal	55,882	50,772	9.14%
Human08	SAMN33744955	Human fecal	80,184	72,324	9.80%
Human09	SAMN33744956	Human fecal	35,824	32,808	8.42%
Human10	SAMN33744957	Human fecal	30,176	27,645	8.39%
Human11	SAMN33744958	Human fecal	78,021	72,165	7.51%
Human12	SAMN33744959	Human fecal	123,564	112,774	8.73%
NHP1	SAMN33744960	Nonhuman primate fecal	37,377	35,637	4.65%
NHP2	SAMN33744961	Nonhuman primate fecal	98,953	92,910	6.11%
NHP3	SAMN33744962	Nonhuman primate fecal	287,932	269,181	6.51%
NHP4	SAMN33744963	Nonhuman primate fecal	56,002	52,080	7.00%
NHP5	SAMN33744964	Nonhuman primate fecal	28,351	26,874	5.21%
NHP6	SAMN33744965	Nonhuman primate fecal	104,900	97,907	6.67%
NHP7	SAMN33744966	Nonhuman primate fecal	28,409	26,415	7.02%
NHP8	SAMN33744967	Nonhuman primate fecal	25,764	23,788	7.67%
NHP9	SAMN33744968	Nonhuman primate fecal	29,434	27,018	8.21%
NHP10	SAMN33744969	Nonhuman primate fecal	58,005	53,206	8.27%
NHP11	SAMN33744970	Nonhuman primate fecal	44,422	39,862	10.26%
NHP12	SAMN33744971	Nonhuman primate fecal	36,887	33,991	7.85%
NHP13	SAMN33744972	Nonhuman primate fecal	55,101	49,958	9.33%
NHP14	SAMN33744973	Nonhuman primate fecal	34,701	31,934	7.97%
NHP15	SAMN33744974	Nonhuman primate fecal	64,954	60,237	7.26%
NHP16	SAMN33744975	Nonhuman primate fecal	50,839	47,371	6.82%
NHP17	SAMN33744976	Nonhuman primate fecal	75,005	68,826	8.24%
NHP18	SAMN33744977	Nonhuman primate fecal	76,770	70,964	7.56%
NHP19	SAMN33744978	Nonhuman primate fecal	46,543	44,239	4.95%

NHP20	SAMN33744979	Nonhuman primate fecal	40,031	37,507	6.31%
NHP21	SAMN33744980	Nonhuman primate fecal	39,344	36,571	7.05%
NHP22	SAMN33744981	Nonhuman primate fecal	29,797	27,118	8.99%
NHP23	SAMN33744982	Nonhuman primate fecal	36,615	33,891	7.44%
NHP24	SAMN33744983	Nonhuman primate fecal	84,056	76,577	8.90%
NHP25	SAMN33744984	Nonhuman primate fecal	27,672	26,198	5.33%
NHP26	SAMN33744985	Nonhuman primate fecal	32,150	28,996	9.81%
NHP27	SAMN33744986	Nonhuman primate fecal	157,483	144,045	8.53%
NHP28	SAMN33744987	Nonhuman primate fecal	31,830	29,320	7.88%
NHP29	SAMN33744988	Nonhuman primate fecal	41,127	37,816	8.05%
NHP30	SAMN33744989	Nonhuman primate fecal	60,491	55,710	7.90%
NHP31	SAMN33744990	Nonhuman primate fecal	74,435	67,968	8.69%
NHP32	SAMN33744991	Nonhuman primate fecal	59,136	54,146	8.44%
NHP33	SAMN33744992	Nonhuman primate fecal	35,473	32,346	8.82%
NHP34	SAMN33744993	Nonhuman primate fecal	39,545	36,508	7.68%
NHP35	SAMN33744994	Nonhuman primate fecal	33,505	31,048	7.33%
NHP36	SAMN33744995	Nonhuman primate fecal	44,082	41,003	6.98%
NHP37	SAMN33744996	Nonhuman primate fecal	59,451	54,867	7.71%
NHP38	SAMN33744997	Nonhuman primate fecal	14,879	13,872	6.77%
NHP39	SAMN33744998	Nonhuman primate fecal	71,275	64,595	9.37%
NHP40	SAMN33744999	Nonhuman primate fecal	62,042	57,199	7.81%

Supplementary Protocol

VESPA Protocol

March 2023

Contents

1) Starting material

Starting material can be fresh, freshly frozen (no buffer), or stored ~1:1 in RNA later.

Sample types tested:

Feces
Intestine- tissue
Entamoeba cysts
Vomit
Intestine- contents
Whole helminths
Stomach- contents
Environmental
Tapeworm proglottids/cysts

2) gDNA extraction

Use

- Qiagen DNeasy PowerLyzer PowerSoil Kit (catalog #12855-5)

according to manufacturer's instructions.

Weigh out up to .20 g of input feces or .25 g of input for all other sample types.

Elute in 100 µl C6 buffer (included in kit) and store at -20 °C.

3) 18S V4 Amplicon PCR

Set up amplicon PCR reactions *in triplicate.*

Use

- Invitrogen Platinum II Hot Start 2X PCR Master Mix (Catalog # 14000012)

with the following reaction and cycling conditions:

Reaction component	Final Conc.	1 x 12.5 µl rxn. (µl)
2X Platinum II HotStart PCR Master Mix*	1X	6.0
10 µM Forward primer	0.2 µM	0.25
10 µM Reverse primer	0.2 µM	0.25
Platinum II GC Enhancer*	NA	2.5
Nuclease-free water*	NA	2.5
~10 ng/µl gDNA	0.8 ng/µl	1.0
		12.5 µl

*Included in Master Mix Kit

Primers **Locus-specific sequence** Nextera adapter sequence

Forward: 29_F
AGCAGCCGCGGTAATTCCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

Reverse: 21b8_I_R

TCCGTCAATTYCTTIAASTTTCGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Step	Temp °C	Time	Cycles
Activation	94	2 min	1
Denaturation	94	15 sec	30
Annealing	60	15 sec	
Extension	68	15 sec	
Final hold	4	hold	

4) Amplicon cleanup

Use

- Beckman Coulter Ampure XP beads (catalog #A63880)

and

- Magnetic particle separator (MPC).

Always make 75% Ethanol immediately prior to use.

1- Shake Ampure XP beads at room temperature for > 30 minutes prior to use.
2- *Pool* all 3 PCR reactions into a single plate or tube and mix by pipetting (~37.5 µl).
3- Remove 7.5 µl and store at -20 °C if you would like to visualize bands on a gel (~30 µl).
4- Add AMPure XP beads for **0.8X RATIO** (*e.g.* 24 µl beads per 30 µl product).
5- Gently pipette up and down 15 times.
6- Incubate at room temperature for 5 minutes.
7- Put tubes on MPC and incubate at room temperature for 2 minutes.
8- Remove and discard supernatant.
9- With tubes on MPC, add 175 µl of 75% ethanol.
10- Wait >1 minute.
11- Remove and discard supernatant.
12- Add 175 µl of 75% ethanol.
13- Wait >1 minute.
14- Remove and discard supernatant.
15- Remove all ethanol with P20 tips.
16- With tubes on MPC, let the pellet air-dry for 5 minutes.
17- Add **47 µl of Tris pH 8.5.**
18- Remove tubes from MPC and gently pipette up and down to resuspend beads.
19- Incubate at room temperature for 2 minutes.
20- Put tubes on MPV and incubate at room temperature for 2 minutes.
21- Carefully transfer **45 µl** of supernatant to a new PCR tubes or plate.

5) Indexing PCR

Set up Indexing PCR reactions *on ice.*

Use

- Roche KAPA HiFi HotStart ReadyMix (catalog #KK2601)

and

- IDT for Illumina Nextera DNA Unique Dual Indexes (catalog #20027215)

with the following reaction and cycling conditions:

Reaction component	1 x 12.5 µl rxn. (µl)
2X KAPA HiFi HotStart ReadyMix	6.0
Nextera Unique Dual Index	2.5
Nuclease-free water	3.0
Clean amplicons in Tris pH 8.5	1.0
	12.5 µl

Step	Temp °C	Time	Cycles
Activation	95 °C	3 min	1
Denaturation	95 °C	30 sec	10
Annealing	55 °C	30 sec	
Extension	72 °C	30 sec	
Final extension	72 °C	5 min	1
Final hold	4 °C	hold	

6) Library cleanup

Use

- Beckman Coulter Ampure XP beads (catalog #A63880)

and

- Magnetic particle separator (MPC).

Always make 75% Ethanol immediately prior to use.

1- Shake Ampure XP beads at room temperature for > 30 minutes prior to use.
2- Add AMPure XP beads for **0.8X RATIO** (*e.g.* 9.6 µl beads per 12.5 µl PCR product).
3- Gently pipette up and down 15 times.
4- Incubate at room temperature for 5 minutes.
5- Put tubes on MPC and incubate at room temperature for 2 minutes.
6- Remove and discard supernatant.
7- With tubes on MPC, add 175 µl of 75% ethanol.
8- Wait >1 minute.
9- Remove and discard supernatant.
10- Add 175 µl of 75% ethanol.
11- Wait >1 minute.
12- Remove and discard supernatant.
13- Remove all ethanol with P20 tips.
14- With tubes on MPC, let the pellet air-dry for 5 minutes.
15- Add **22 µl of Tris pH 8.5.**
16- Remove tubes from MPC and gently pipette up and down to resuspend beads.
17- Incubate at room temperature for 2 minutes.
18- Put tubes on MPV and incubate at room temperature for 2 minutes.
19- Carefully transfer **20 µl** of supernatant to a new PCR tubes or plate.

7) Quantification and size determination

Use

- Invitrogen Qubit Fluorimeter and dsDNA High-Sensitivity Assay Kit (catalog #Q33230)

and

- Agilent Bioanalyzer and Agilent High Sensitivity DNA Kit (catalog #5067-4626)

according to manufacturer’s instructions.

Measure the concentration of each library using a Qubit fluorometer and 3 µl of each library.

Measure the size of each library or a representative subset of libraries using an Agilent Bioanalyzer and 1 µl of a 1 ng/µl dilution (in Tris pH 8.5) of the library for a total of 1 ng.

8) Pooling and sequencing

Requirements for core facility submission/in-house sequencing will determine pooling specifics. Run on an Illumina MiSeq instrument, 300 x 300 cycle chemistry, and add 10 – 20% PhiX.

Supplemental Literature Review Table 1. All Literature review results

Article type	Title	Authors	Journal/Book	Year
Research	Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity	Porazinska DL, GIBLIN-DAVIS RM, Faller L, Farmerie W, Kanzaki N, Morris K, Powers TO, Tucker AE, Sung WA, Thomas WK	Molecular ecology resources	2009
Review	Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions	Wegener Parfrey L, Walters WA, Knight R	Frontiers in microbiology	2011
Research	A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples	Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, Hoffmann C, Nabel CS, Hill DA, Artis D, Bachman MA, Custers-Allen R	Genome biology	2012
Research	PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences	Machida RJ, Knowlton N	PLoS one	2012
Research	A megafauna's microfauna: gastrointestinal parasites of New Zealand's extinct moa (Aves: Dinornithiformes)	Wood JR, Wilmshurst JM, Rawlence NJ, Bonner KI, Worthy TH, Kinsella JM, Cooper A	PloS one	2013
Research	Assessment of helminth biodiversity in wild rats using 18S rDNA based metagenomics	Tanaka R, Hino A, Tsai IJ, et al	PloS one	2014
Research	Characterization of the 18S rRNA gene for designing universal eukaryote specific primers	Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C	PloS one	2014
Research	Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity	Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, Kodira C, Mohiuddin M, Brunelle J, Driscoll M, Fierer N	Frontiers in microbiology	2014
Research	Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia	Hugerth LW, Muller EE, Hu YO, Lebrun LA, Roume H, Lundin D, Wilmes P, Andersson AF	PloS one	2014
Review	Are human intestinal eukaryotes beneficial or commensals?	Lukeš J, Stensvold CR, Jirků-Pomajbíková K, Parfrey LW	PLoS pathogens	2015

Research	Exploring the gastrointestinal “nemabiome”: deep amplicon sequencing to quantify the species composition of parasitic nematode communities	Avramenko RW, Redman EM, Lewis R, Yazwinski TA, Wasmuth JD, Gilleard JS	PLoS One	2015
Research	Parasitic nematode communities of the red kangaroo, Macropus rufus: richness and structuring in captive systems	Lott, M. J., Hose, G. C. and Power, M. L	Parasitology Research	2015
Research	Tracking year-to-year changes in intestinal nematode communities of rufous mouse lemurs (Microcebus rufus)	Aivelo T, Medlar A, LÃ¶ytynoja A, Laakkonen J, Jernvall J.	Parasitology	2015
Research	A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method	Hino A, Maruyama H, Kikuchi T	Parasitology international	2016
Research	The utility of diversity profiling using Illumina 18S rRNA gene amplicon deep sequencing to detect and discriminate Toxoplasma gondii among the cyst-forming coccidia	Cooper MK, Phalen DN, Donahoe SL, Rose K, Šlapeta J	Veterinary parasitology	2016
Research	Deep-sequencing to resolve complex diversity of apicomplexan parasites in platypuses and echidnas: Proof of principle for wildlife disease investigation.	Šlapeta J, Saverimuttu S, Vogelnest L, Sangster C, Hulst F, Rose K, Thompson P, Whittington R	Infection, Genetics and Evolution	2017
Methods	Microbiome Helper: a Custom and Streamlined Workflow for MIcrobiome Research	Comeau AM, Douglas GM, Langille MG	MSystems	2017
Review	Mosquito vector-associated microbiota: Metabarcoding bacteria and eukaryotic symbionts across habitat types in Thailand endemic for dengue and other arthropod-borne diseases	Thongsripong P, Chandler JA, Green AB, Kittayapong P, Wilcox BA, Kapan DD, Bennett SN	Ecology and evolution	2017
Review	Omic” investigations of protozoa and worms for a deeper understanding of the human gut “parasitome	Marzano V, Mancinelli L, Bracaglia G, Del Chierico F, Vernocchi P, Di Girolamo F, Garrone S, Tchidjou Kuekou H, D’Argenio P, Dallapiccola B, Urbani A	PLoS neglected tropical diseases	2017
Research	Performance of DNA metabarcoding, standard barcoding, and morphological approach in the identification of host-parasitoid interactions	Šigut M, Kostovčík M, Šigutová H, Hulcr J, Drozd P, Hrček J	PLoS One	2017
Research	Preferential suppression of Anopheles gambiae host sequences allows detection of the mosquito eukaryotic microbiome	Belda E, Coulibaly B, Fofana A, Beavogui AH, Traore SF, Gohl DM, Vernick KD, Riehle MM	Scientific reports	2017
Research	Small subunit ribosomal metabarcoding reveals extraordinary trypanosomatid diversity in Brazilian bats	Dario MA, da Rocha RM, Schwabl P, Jansen AM, Llewellyn MS	PLoS neglected tropical diseases	2017

Research	The use of nemabiome metabarcoding to explore gastro-intestinal nematode species diversity and anthelmintic treatment effectiveness in beef calves	Avramenko RW, Redman EM, Lewis R, Bichuette MA, Palmeira BM, Yazwinski TA, Gilleard JS.	International journal for parasitology	2017
Research	Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome	Krogsgaard LR, Andersen LO', Johannesen TB, et al.	Clinical and translational gastroenterology	2018
Research	Characterization of ecto- and endoparasite communities of wild Mediterranean teleosts by a metabarcoding approach	Scheifler M, Ruiz-Rodríguez M, Sanchez-Brosseau S, Magnanou E, Suzuki MT, West N, Duperron S, Desdevises Y	PloS one	2018
Research	Coprolites reveal ecological interactions lost with the extinction of New Zealand birds	Boast AP, Weyrich LS, Wood JR, Metcalf JL, Knight R, Cooper A.	Proceedings of the National Academy of Sciences	2018
Research	Deep sequencing reveals multiclonality and new discrete typing units of Trypanosoma cruzi in rodents from the southern United States	Pronovost H, Peterson AC, Chavez BG, Blum MJ, Dumonteil E, Herrera CP	Journal of Microbiology, Immunology and Infection	2018
Research	Diversity of Entamoeba spp. in African great apes and humans: an insight from Illumina MiSeq high-throughput sequencing	Vlčková K, Kreisinger J, Pafčo B, Čížková D, Tagg N, Hehl AB, Modrý D	International journal for parasitology	2018
Review	DNA barcoding of ancient parasites	Wood JR.	Parasitology	2018
Research	Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys	Grey EK, Bernatchez L, Cassey P, Deiner K, Deveney M, Howland KL, Lacoursière-Roussel A, Leong SC, Li Y, Olds B, Pfrender ME	Scientific Reports	2018
Research	Eukaryotes in the gut microbiota in myalgic encephalomyelitis/chronic fatigue syndrome	Mandarano AH, Giloteaux L, Keller BA, Levine SM, Hanson MR	PeerJ	2018
Research	Exploring non-invasive sampling of parasites by metabarcoding gastrointestinal nematodes in Madagascar frog species.	Aivelo T, Harris K, Cadle JE, Wright P	Basic and Applied Herpetology	2018
Research	High species diversity of trichostrongyle parasite communities within and between Western Canadian commercial and conservation bison herds revealed by nemabiome metabarcoding	Avramenko RW, Bras A, Redman EM, Woodbury MR, Wagner B, Shury T, Liccioli S, Windeyer MC, Gilleard JS.	Parasites & vectors	2018
Review	Meta-taxonomic analysis of prokaryotic and eukaryotic gut flora in stool samples from visceral leishmaniasis cases and endemic controls in Bihar State India	Hamad I, Abou Abdallah R, Ravaux I, Mokhtari S, Tissot-Dupont H, Michelle C, Stein	PLoS One	2018

		A, Lagier JC, Raoult D, Bittar F		
Research	Metabarcoding analysis of eukaryotic microbiota in the gut of HIV-infected patients.	Pafčo B, Čížková D, Kreisinger J, Hasegawa H, Vallo P, Shutt K, Todd A, Petrželková KJ, Modrý D	Scientific reports	2018
Research	Metabarcoding Fecal DNA Reveals Extent of Halichoerus grypus (Gray Seal) Foraging on Invertebrates and Incidence of Parasite Exposure	Aivelo T, Medlar A, Löytynoja A, Laakkonen J, Jernvall J	International Journal of Primatology	2018
Review	Microbial eukaryotes: a missing link in gut microbiome studies	Laforest-Lapointe I, Arrieta MC	MSystems	2018
Research	Multifaceted DNA metabarcoding: Validation of a noninvasive, next-generation approach to studying bat populations	Swift JF, Lance RF, Guan X, Britzke ER, Lindsay DL, Edwards CE.	Evolutionary applications	2018
Research	Museum metabarcoding: A novel method revealing gut helminth communities of small mammals across space and time	Greiman SE, Cook JA, Tkach VV, Hoberg EP, Menning DM, Hope AG, Sonsthagen SA, Talbot SL	International journal for parasitology	2018
Research	New Determination of Prey and Parasite Species for Northern Indian Ocean Blue Whales	de Vos A, Faux CE, Marthick J, Dickinson J, Jarman SN	Frontiers in Marine Science	2018
Review	Opportunities and challenges in metabarcoding approaches for helminth community identification in wild mammals	Aivelo T, Medlar A.	Parasitology	2018
Research	Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in blood via targeted amplicon deep sequencing	Flaherty BR, Talundzic E, Barratt J, Kines KJ, Olsen C, Lane M, Sheth M, Bradbury RS	Microbiome	2018
Research	Scrutinizing key steps for reliable metabarcoding of environmental samples	Alberdi A, Aizpurua O, Gilbert MT, Bohmann K	Methods in Ecology and Evolution.	2018
Research	Unprecedented Symbiont Eukaryote Diversity Is Governed by Internal Trophic Webs in a Wild Non-Human Primate	Wilcox JJ, Hollocher H	Protist	2018
Research	Ancient parasite DNA from late Quaternary Atacama Desert rodent middens	Wood JR, Díaz FP, Latorre C, Wilmshurst JM, Burge OR, González F, Gutiérrez RA	Quaternary Science Reviews	2019
Research	Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline	Pauvert C, Buée M, Laval V, Edel-Hermann V, Fauchery L, Gautier A, Lesur I, Vallance J, Vacher C	Fungal Ecology	2019
Research	Genetic diversity of primate strongylid nematodes: Do sympatric nonhuman primates and humans share their strongylid worms?	Pafčo B, Kreisinger J, Čížková D, Pšenková-Profousová I,	Molecular ecology	2019

		Shutt-Phillips K, Todd A, Fuh T, Petrželková KJ, Modrý D		
Review	High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations	Tedersoo L, Drenkhan R, Anslan S, Morales-Rodriguez C, Cleary M	Molecular ecology resources	2019
Review	High-Throughput Sequencing for Understanding the Ecology of Emerging Infectious Diseases at the Wildlife-Human Interface	Titcomb GC, Jerde CL, Young HS	Frontiers in Ecology and Evolution	2019
Research	How quantitative is metabarcoding: A meta-analytical approach	Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI	Molecular Ecology	2019
Research	Metabarcoding of eukaryotic parasite communities describes diverse parasite assemblages spanning the primate phylogeny	Holmes IA, Monagan Jr IV, Rabosky DL, Davis Rabosky AR	Ecology and evolution	2019
Research	Metagenomics and microscope revealed T. trichiura and other intestinal parasites in a cesspit of an Italian nineteenth century aristocratic palace	Lappan R, Classon C, Kumar S, Singh OP, De Almeida RV, Chakravarty J, Kumari P, Kansal S, Sundar S, Blackwell JM	PLoS neglected tropical diseases	2019
Research	The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative	Piñol J, Senar MA, Symondson WO	Molecular ecology	2019
Research	Validation of a universal set of primers to study animal-associated microeukaryotic communities	Del Campo J, Pons MJ, Herranz M, Wakeman KC, Del Valle J, Vermeij MJ, Leander BS, Keeling PJ	Environmental microbiology	2019
Review	A global parasite conservation plan	Carlson CJ, Hopkins S, Bell KC, Doña J, Godfrey SS, Kwak ML, Lafferty KD, Moir ML, Speer KA, Strona G, Torchin M	Biological Conservation	2020
Research	A metabarcoding approach detects rare blood parasites in fossorial amphisbaenians	Harris DJ, Pereira A, Perera A	North-Western Journal of Zoology	2020
Research	A new method of metabarcoding Microsporidia and their hosts reveals high levels of microsporidian infections in mosquitoes (Culicidae)	Trzebny A, Slodkowicz-Kowalska A, Becnel JJ, Sanscrainte N, Dabert M	Molecular Ecology Resources	2020
Research	A novel metabarcoded 18S ribosomal DNA sequencing tool for the detection of Plasmodium species in malaria positive patients	Wahab A, Shaukat A, Ali Q, Hussain M, Khan TA, Khan MA, Rashid I, Saleem MA,	Infection, Genetics and Evolution	2020

		Evans M, Sargison ND, Chaudhry U		
Review	All together now: Limitations and recommendations for the simultaneous analysis of all eukaryotic soil sequences	Jurburg SD, Keil P, Singh BK, Chase JM	Molecular Ecology Resources	2020
Research	Alternative primers to identify a range of apicomplexan parasites	Pinheiro SN, de Souza MF, Oliveira CB, Neto VF, Lanza DC	Journal of microbiological methods	2020
Research	Application of next generation sequencing for detection of protozoan pathogens in shellfish	DeMone C, Hwang MH, Feng Z, McClure JT, Greenwood SJ, Fung R, Kim M, Weese JS, Shapiro K	Food and waterborne parasitology	2020
Research	Biases in bulk: DNA metabarcoding of marine communities and the methodology involved	van der Loos LM, Nijland R	Molecular Ecology	2020
Methods	BIOCOM-PIPE: a new user-friendly metabarcoding pipeline for the characterization of microbial diversity from 16S, 18S and 23S rRNA gene amplicons	Djemiel C, Dequiedt S, Karimi B, Cottin A, Girier T, El Djoudi Y, Wincker P, Lelièvre M, Mondy S, Prévost-Bouré NC, Maron PA	BMC bioinformatics	2020
Research	Biodiversity of protists and nematodes in the wild nonhuman primate gut	Mann AE, Mazel F, Lemay MA, Morien E, Billy V, Kowalewski M, Di Fiore A, Link A, Goldberg TL, Tecot S, Baden AL	The ISME journal	2020
Research	Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes	Giebner H, Langen K, Bourlat SJ, Kukowka S, Mayer C, Astrin JJ, Misof B, Fonseca VG	Molecular Ecology Resources	2020
Review	Contrasting strategies: human eukaryotic versus bacterial microbiome research	Hooks KB, O'Malley MA	Journal of Eukaryotic Microbiology	2020
Research	Evaluation of Metabarcoding Primers for Analysis of Soil Nematode Communities	Sikder M, Vestergård M, Sapkota R, Kyndt T, Nicolaisen M	Diversity	2020
Research	Evidence of Batrachochytrium dendrobatidis and other amphibian parasites in the Green toad (Bufotes viridis), syntopic amphibians and environment in the Cologne Bay, Germany	Sachs M, Schluckebier R, Poll K, Schulz V, Sabino-Pinto J, Schmidt E, Simon K, Kuenzel S, Ziegler T, Arndt H, Vences M	SALAMANDRA	2020
Research	Exploring micro-eukaryotic diversity in the gut: Co-occurrence of Blastocystis subtypes and other protists in zoo animals	Betts EL, Gentekaki E, Tsaousis AD.	Frontiers in microbiology	2020

Research	Exploring Prokaryotic and Eukaryotic Microbiomes Helps in Detecting Tick-Borne Infectious Agents in the Blood of Camels	Mohamed WM, Ali AO, Mahmoud HY, Omar MA, Chatanga E, Salim B, Naguib D, Anders JL, Nonaka N, Moustafa MA, Nakao R	Pathogens	2020
Research	Genetic diversity and multiplicity of infection in Fasciola gigantica isolates of Pakistani livestock	Rehman ZU, Zahid O, Rashid I, Ali Q, Akbar MH, Oneeb M, Shehzad W, Ashraf K, Sargison ND, Chaudhry U	Parasitology international	2020
Research	Genetic diversity of Trypanosoma cruzi parasites infecting dogs in southern Louisiana sheds light on parasite transmission cycles and serological diagnostic performance	Dumonteil E, Elmayan A, Majeau A, Tu W, Duhon B, Marx P, Wolfson W, Balsamo G, Herrera C	PLOS Neglected Tropical Diseases	2020
Research	Gut eukaryotic communities in pigs: diversity, composition and host genetics contribution	Ramayo-Caldas Y, Prenafeta-Boldú F, Zingaretti LM, Gonzalez-Rodriguez O, Dalmau A, Quintanilla R, Ballester M	Animal Microbiome	2020
Research	High levels of third-stage larvae (L3) overwinter survival for multiple cattle gastrointestinal nematode species on western Canadian pastures as revealed by ITS2 rDNA metabarcoding.	Wang T, Avramenko RW, Redman EM, Wit J, Gilleard JS, Colwell DD.	Parasites & vectors	2020
Research	Humic-acid-driven escape from eye parasites revealed by RNA-seq and target-specific metabarcoding	Noreikiene K, Ozerov M, Ahmad F, Kõiv T, Kahar S, Gross R, Sepp M, Pellizzone A, Vesterinen EJ, Kisand V, Vasemägi A	Parasites & vectors	2020
Research	Investigation of the piroplasm diversity circulating in wildlife and cattle of the greater Kafue ecosystem, Zambia	Squarre D, Nakamura Y, Hayashida K, Kawai N, Chambaro H, Namangala B, Sugimoto C, Yamagishi J	Parasites & Vectors	2020
Methods	MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding	Arranz V, Pearman WS, Aguirre JD, Liggins	Scientific Data	2020
Research	Metabarcoding analysis of strongylid nematode diversity in two sympatric primate species	McCosker C, Flanders K, Ono K, Dufault M, Mellone D, Olson Z	Northeastern Naturalist	2020
Research	Metabarcoding Gastrointestinal Nematodes in Sympatric Endemic and Nonendemic Species in Ranomafana National Park, Madagascar	Gogarten JF, Calvignac-Spencer S, Nunn CL, Ulrich M, Saiepour N, Nielsen HV, Deschner T, Fichtel C,	Molecular ecology resources	2020

		Kappeler PM, Knauf S, Müller-Klein N		
Research	Metabolically similar cohorts of bacteria exhibit strong cooccurrence patterns with diet items and eukaryotic microbes in lizard guts	Chessa D, Murgia M, Sias E, Deligios M, Mazzarello V, Fiamma M, Rovina D, Carenti G, Ganau G, Pintore E, Fiori M	Scientific reports	2020
Review	Microbiome definition re-visited: old concepts and new challenges	Berg G, Rybakova D, Fischer D, Cernava T, Vergès MC, Charles T, Chen X, Cocolin L, Eversole K, Corral GH, Kazou M	Microbiome	2020
Review	Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity	Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, Morton L, Jarman RG	The Journal of infectious diseases	2020
Research	Obtaining deeper insights into microbiome diversity using a simple method to block host and non-targets in amplicon sequencing	Mayer T, Mari A, Almario J, Murillo-Roos M, Abdullah M, Dombrowski N, Hacquard S, Kemen EM, Agler MT	bioRxiv	2020
Methods	PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes	Zafeiropoulos H, Viet HQ, Vasileiadou K, Potirakis A, Arvanitidis C, Topalis P, Pavloudi C, Pafilis E	GigaScience	2020
Review	Post-genomic progress in helminth parasitology	McVeigh P.	Parasitology	2020
Research	Simultaneous analysis of the intestinal parasites and diet through eDNA metabarcoding.	Cabodevilla X, Gómez-Moliner BJ, Madeira MJ	Authorea Preprints	2020
Research	The impact of intragenomic rRNA variation on metabarcoding-derived diversity estimates: A case study from marine nematodes	Pereira TJ, De Santiago A, Schuelke T, Hardy SM, Bik HM	Environmental DNA	2020
Research	The use of ITS-2 rDNA nemabiome metabarcoding to enhance anthelmintic resistance diagnosis and surveillance of ovine gastrointestinal nematodes	Queiroz C, Levy M, Avramenko R, Redman E, Kearns K, Swain L, Silas H, Uehlinger F, Gilleard JS.	International Journal for Parasitology: Drugs and Drug Resistance	2020
Research	Unveiling protist diversity associated with the Pacific oysterCrassostrea gigasusing blocking and excluding primers	Clerissi C, Guillou L, Escoubas JM, Toulza E	BMC microbiology	2020
Research	A repeatable and quantitative DNA metabarcoding assay to characterize mixed strongyle infections in horses	Poissant J, Gavriliuc S, Bellaw J, Redman EM, Avramenko	International Journal for Parasitology	2021

		RW, Robinson D, Workentine ML, Shury TK, Jenkins EJ, McLoughlin PD, Nielsen MK		
Research	Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding	Brandt MI, Trouche B, Quintric L, Günther B, Wincker P, Poulain J, Arnaud-Haond S	Molecular Ecology Resources	2021
Research	Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S	Yeh YC, McNichol JC, Needham DM, Fichot EB, Berdjeb L, Fuhrman JA	Environmental Microbiology	2021
Research	Detection and Identification of Acanthamoeba and Other Nonviral Causes of Infectious Keratitis in Corneal Scrapings by Real-Time PCR and Next-Generation Sequencing-Based 16S-18S Gene Analysis.	Holmgaard DB, Barnadas C, Mirbarati SH, Andersen LO, Nielsen HV, Stensvold CR.	Journal of Clinical Microbiology.	2021
Research	Sensitive universal detection of blood parasites by selective pathogen-DNA enrichment and deep amplicon sequencing	Flaherty BR, Barratt J, Lane M, Talundzic E, Bradbury RS.	Microbiome	2021
Research	Sheep nemabiome diversity and its response to anthelmintic treatment in Swedish sheep herds	Halvarsson P, Höglund J.	Parasites & vectors	2021
Research	The use of CRISPR-Cas Selective Amplicon Sequencing (CCSAS) to reveal the eukaryotic microbiome of metazoans	Zhong KX, Cho A, Deeg CM, Chan AM, Suttle CA.	bioRxiv	2021

Supplemental Literature Review Table 2. Eukaryotic endosymbiont metabarcoding studies

Category	Title	Journal	Year	Marker Gene/s	18S region/s from paper	18S region primer mapping
Pan-parasite/commensal	A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples	Genome biology	2012	18S	not stated	V2
Pan-parasite/commensal	Assessment of helminth biodiversity in wild rats using 18S rDNA based metagenomics	PLoS One	2014	18S	V9	
Nematode-Specific	Tracking year-to-year changes in intestinal nematode communities of rufous mouse lemurs (Microcebus rufus)	Parasitology	2015	18S	not stated	V5-V6
Nematode-Specific	Exploring the gastrointestinal “nemabiome”: deep amplicon sequencing to quantify the species composition of parasitic nematode communities	PLoS One	2015	ITS2		
Pan-parasite/commensal	Parasitic nematode communities of the red kangaroo, Macropus rufus: richness and structuring in captive systems	Parasitology Research	2015	ITS2		
Pan-parasite/commensal	A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method	Parasitology International	2016	18S	V9	
Toxoplasma-specific	The utility of diversity profiling using Illumina 18S rRNA gene amplicon deep sequencing to detect and discriminate Toxoplasma gondii among the cyst-forming coccidia	Veterinary parasitology	2016	18S	V1-V3, V9	
Pan-parasite/commensal	Deep-sequencing to resolve complex diversity of apicomplexan parasites in platypuses and echidnas: Proof of principle for wildlife disease investigation.	Infection, Genetics and Evolution	2017	18S	not stated	V2-V3

Pan-parasite/commensal	Preferential suppression of Anopheles gambiae host sequences allows detection of the mosquito eukaryotic microbiome	Scientific reports	2017	18S	V4, V9	
Trypanosome-specific	Small subunit ribosomal metabarcoding reveals extraordinary trypanosomatid diversity in Brazilian bats	PLoS neglected tropical diseases	2017	18S	V4-V5	
Nematode-Specific	The use of nemabiome metabarcoding to explore gastro-intestinal nematode species diversity and anthelmintic treatment effectiveness in beef calves	International journal for parasitology	2017	ITS2		
Nematode-Specific	Museum metabarcoding: A novel method revealing gut helminth communities of small mammals across space and time	International journal for parasitology	2018	12S, 16S, 23S		
Entamoeba-specific	Diversity of Entamoeba spp. in African great apes and humans: an insight from Illumina MiSeq high-throughput sequencing	International journal for parasitology	2018	18S	not stated	V5
Nematode-Specific	Exploring non-invasive sampling of parasites by metabarcoding gastrointestinal nematodes in Madagascar frog species.	Basic and Applied Herpetology	2018	18S	not stated	V5-V6
Nematode-Specific	Metabarcoding Gastrointestinal Nematodes in Sympatric Endemic and Nonendemic Species in Ranomafana National Park, Madagascar	International Journal of Primatology	2018	18S	not stated	V5-V6
Pan-parasite/commensal	Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome	Clinical and Translational Gastroenterolog y	2018	18S	V4	
Pan-parasite/commensal	Coprolites reveal ecological interactions lost with the extinction of New Zealand birds	Proceedings of the National Academy of Sciences	2018	18S	V9	
Pan-parasite/commensal	Eukaryotes in the gut microbiota in myalgic encephalomyelitis/chronic fatigue syndrome	PeerJ	2018	18S	V9	

Pan-parasite/commensal	New Determination of Prey and Parasite Species for Northern Indian Ocean Blue Whales	Frontiers in Marine Science	2018	18S	not stated	V9
Pan-parasite/commensal	Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in blood via targeted amplicon deep sequencing	Microbiome	2018	18S	not stated	V3
Pan-parasite/commensal	Unprecedented Symbiont Eukaryote Diversity Is Governed by Internal Trophic Webs in a Wild Non-Human Primate	Protist	2018	18S	V9	
Nematode-Specific	High species diversity of trichostrongyle parasite communities within and between Western Canadian commercial and conservation bison herds revealed by nemabiome metabarcoding	Parasites & vectors	2018	ITS2		
Nematode-Specific	Metabarcoding analysis of strongylid nematode diversity in two sympatric primate species	Scientific reports	2018	ITS2		
Trypanosome-specific	Deep sequencing reveals multiclonality and new discrete typing units of Trypanosoma cruzi in rodents from the southern United States	Journal of Microbiology, Immunology and Infection	2018	Tcl DTU (mini-exon)		
Pan-parasite/commensal	Ancient parasite DNA from late Quaternary Atacama Desert rodent middens	Quaternary Science Reviews	2019	18S	V4	
Pan-parasite/commensal	Assessing the Diversity and Distribution of Apicomplexans in Host and Free-Living Environments Using High-Throughput Amplicon Data and a Phylogenetically Informed Reference Framework	Fronteirs in microbiology	2019	18S	V4, V9	
Pan-parasite/commensal	Meta-taxonomic analysis of prokaryotic and eukaryotic gut flora in stool samples from visceral leishmaniasis cases and endemic controls in Bihar State India	PLoS neglected tropical diseases	2019	18S	V9	
Pan-parasite/commensal	Metabolically similar cohorts of bacteria exhibit strong cooccurrence patterns with diet items and eukaryotic microbes in lizard guts	Ecology and evolution	2019	18S	not stated	V2

Pan-parasite/commensal	Validation of a universal set of primers to study animal-associated microeukaryotic communities	Environmental microbiology	2019	18S	V4	
Pan-parasite/commensal	Characterization of ecto- and endoparasite communities of wild Mediterranean teleosts by a metabarcoding approach	PLoS One	2019	Bacterial 16S (incidental reads)		
Nematode-Specific	Genetic diversity of primate strongylid nematodes: Do sympatric nonhuman primates and humans share their strongylid worms?	Molecular ecology	2019	ITS2		
Apicomplexan/Piroplasm/Plasmodium-specific	A novel metabarcoded 18S ribosomal DNA sequencing tool for the detection of Plasmodium species in malaria positive patients	Infection, Genetics and Evolution	2020	18S	not stated	V4
Apicomplexan/Piroplasm/Plasmodium-specific	Investigation of the piroplasm diversity circulating in wildlife and cattle of the greater Kafue ecosystem, Zambia	Parasites & Vectors	2020	18S	not stated	V4
Pan-parasite/commensal	A metabarcoding approach detects rare blood parasites in fossorial amphisbaenians	North-Western Journal of zoology	2020	18S	not stated	V9
Pan-parasite/commensal	Biodiversity of protists and nematodes in the wild nonhuman primate gut	The ISME journal	2020	18S	V4	
Pan-parasite/commensal	Evidence of Batrachochytrium dendrobatidis and other amphibian parasites in the Green toad (Bufotes viridis), syntopic amphibians and environment in the Cologne Bay, Germany	SALAMANDRA	2020	18S	V9	
Pan-parasite/commensal	Exploring Prokaryotic and Eukaryotic Microbiomes Helps in Detecting Tick-Borne Infectious Agents in the Blood of Camels	Pathogens	2020	18S	V4	
Pan-parasite/commensal	Metabarcoding of eukaryotic parasite communities describes diverse parasite assemblages spanning the primate phylogeny	Molecular ecology resources	2020	18S	V4	

Pan-parasite/commensal	Metagenomics and microscope revealed T. trichiura and other intestinal parasites in a cesspit of an Italian nineteenth century aristocratic palace	Scientific reports	2020	18S	not stated	V2
Pan-parasite/commensal	Parasites of an Arctic scavenger; the wolverine (Gulo gulo)	International Journal for Parasitology: Parasites and Wildlife	2020	18S	V9	
Pan-parasite/commensal	Simultaneous analysis of the intestinal parasites and diet through eDNA metabarcoding.	Authorea Preprints	2020	18S	V3, V7-V8	
Pan-parasite/commensal	Metabarcoding Fecal DNA Reveals Extent of Halichoerus grypus (Gray Seal) Foraging on Invertebrates and Incidence of Parasite Exposure	Northeastern Naturalist	2020	18S	not stated	V5
Protozoal genera-specific	Exploring micro-eukaryotic diversity in the gut: Co-occurrence of Blastocystis subtypes and other protists in zoo animals	Fronteirs in microbiology	2020	18S	not stated	
Nematode-Specific	Evaluation of Metabarcoding Primers for Analysis of Soil Nematode Communities	Diversity	2020	18S, COI	V1-V2, V4, V6-V8	
Pan-parasite/commensal	Gut eukaryotic communities in pigs: diversity, composition and host genetics contribution	Animal Microbiome	2020	18S, ITS2	V4	
Diplostomidae-specific	Humic-acid-driven escape from eye parasites revealed by RNA-seq and target-specific metabarcoding	Parasites & vectors	2020	COI		
Nematode-Specific	Assessing anthelmintic resistance risk in the post-genomic era: a proof-of-concept study assessing the potential for widespread benzimidazole-resistant gastrointestinal nematodes in North American cattle and bison	Parasitology	2020	ITS2		
Nematode-Specific	High levels of third-stage larvae (L3) overwinter survival for multiple cattle gastrointestinal nematode species on western Canadian pastures as revealed by ITS2 rDNA metabarcoding.	Parasites & vectors	2020	ITS2		

Nematode-Specific	The use of ITS-2 rDNA nemabiome metabarcoding to enhance anthelmintic resistance diagnosis and surveillance of ovine gastrointestinal nematodes	International Journal for Parasitology: Drugs and Drug Resistance	2020	ITS2		
Trematode-specific	Genetic diversity and multiplicity of infection in Fasciola gigantica isolates of Pakistani livestock	Parasitology international	2020	ITS2, nt-MD1		
Trypanosome-specific	Genetic diversity of Trypanosoma cruzi parasites infecting dogs in southern Louisiana sheds light on parasite transmission cycles and serological diagnostic performance	PLOS Neglected Tropical Diseases	2020	Tcl DTU (mini-exon)		
Pan-parasite/commensal	Sensitive universal detection of blood parasites by selective pathogen-DNA enrichment and deep amplicon sequencing	Microbiome	2021	18S	not stated	V3
Nematode-Specific	A repeatable and quantitative DNA metabarcoding assay to characterize mixed strongyle infections in horses	International Journal for Parasitology	2021	ITS2		
Nematode-Specific	Sheep nemabiome diversity and its response to anthelmintic treatment in Swedish sheep herds	Parasites & vectors	2021	ITS2		

Supplemental Literature Review Table 3. Broad 18S Primers

Set ID	Used in (only metabarcoding studies listed)*	Primers designed to target	Primers designed to avoid	Gene	Locus	FWD primer reference	FWD primer name	REV primer reference	R primer name	Size (based on *E. histolytica* mapping)
Bates 2012	Bates 2012*, Parfrey 2014	all eukaryotes		18S	V4-V5	Bates 2012	515f	Bates 2012	1119r	600
Bower 2004	Del Campo 2019b, Clerissi 2020, Mohamed 2020	all protozoans	metazoans	18S	V4-V5	Bower 2004	18S-EUK581-F	Bower 2004	18S-EUK1134-R	550
Bradley 2016	Bradley 2016*	all eukaryotes	primer bias (no degeneracy)	18S	V4	Stoeck 2010	TAReuk454FWD1	Bradley 2016	V4r	375
Cavalier-Smith 2009/Brate 2010- 1	Lohan 2016*	all eukaryotes		18S	V4	Cavalier-Smith 2009	3NDf	Brate 2010	V4_euk_R1	450
Cavalier-Smith 2009/Brate 2010-2	Lohan 2016*	all eukaryotes		18S	V4	Cavalier-Smith 2009	3NDf	Brate 2010	V4_euk_R2	450
Cavalier-Smith 2009/Geisen 2018	Geisen 2018*	all eukaryotes		18S	V4-V5	Cavalier-Smith 2009	3NDf	Geisen 2018	1132mod	625
Comeau 2011	Comeau 2011*, del Campo 2019, Mann 2020, Mohamed 2020	all eukaryotes	prokaryotes	18S	V4	Comeau 2011	E572F	Comeau 2011	E1009R	500
DeMone 2020	DeMone 2020	all protozoans		18S	V4	DeMone 2020	18SV4_F	DeMone 2020	18SV4_R	375
DeMone 2020-Giardia	DeMone 2020	Giardia		18S	V4	DeMone 2020	18SV4_F	DeMone 2020	Giardia 18SV4_R	375
Hadziavdic 2014- 566	Cabodevilla 2020, Mayer 2020, Ramayo-Caldas 2020, Hadziavdic 2014	all eukaryotes	prokaryotes	18S	V4-V5	Hadziavdic 2014	F-566	Hadziavdic 2014	R-1200	475

Hadziavdic 2014- 574	Grey 2018*	all eukaryotes	prokaryotes, primer bias (no degeneracy)	18S	V4	Hadziavdic 2014	F-574	Hadziavdic 2014	R-952	650
Hugerth 2014-563	Hugerth 2014	all eukaryotes	prokaryotes, host	18S	V4-V5	Hugerth 2014	616*	Hugerth 2014	1132	500
Hugerth 2014-574	Hugerth 2014	all eukaryotes	prokaryotes, host	18S	V4-V5	Hugerth 2014	574*	Hugerth 2014	1132	550
Hugerth 2014-616	Hugerth 2014	all eukaryotes	prokaryotes, host	18S	V4-V5	Hugerth 2014	563	Hugerth 2014	1132	570
Krogsgaard 2018- G3	Krogsgaard 2018, Gogarten 2020	all eukaryotes		18S	V4-V5	Krogsgaard 2018	G4F3	Krogsgaard 2018	G4R3	380
Krogsgaard 2018- G4	Krogsgaard 2018, Gogarten 2020	all eukaryotes		18S	V4	Krogsgaard 2018	G3F1	Krogsgaard 2018	G3R1	450
Krogsgaard 2018- G6	Krogsgaard 2018, Gogarten 2020	all eukaryotes		18S	V4	Krogsgaard 2018	G6F1	Krogsgaard 2018	G6R1	575
Machida 2012	Machida 2012	all metazoans		18S	V4-V5	Machida 2012	18S#1	Machida 2012	18S#2RC	600
Sikder 2020	Sikder 2020	nematodes		18S	V4-V5	Sikder 2020	MMSF	Sikder 2020	MMSR	600
Stoeck 2010	Stoeck 2010*, Belda 2017, Del Campo 2019a, Clerissi 2020	all eukaryotes including environmental clones		18S	V4	Stoeck 2010	TAReuk454FWD1	Stoeck 2010	TAReukREV3	400
Wood 2013	Wood 2013	broad range of invertebrates		18S	V4	Wood 2013	Nem18SlongF	Wood 2013	Nem18SlongR	425

*Environmental metabarcoding study

CHAPTER 5: Conclusion

Summary

Overall, the work in this thesis concerns the evaluation of existing methodologies to study eukaryotic endosymbiont communities, the development of a new and improved methodology for this purpose, and its application to clinical samples to assess its "real world" performance. To this end, I used sequencing-based methods to discover a new bacterial species and disease-associated pathogen in **Chapter 2.** Part of this investigation involved an attempt to characterize eukaryotic parasites using published metabarcoding methods, revealing biases and deficiencies in those methods. **Chapter 3** addressed the concern of an overabundance of interfering host sequence reads with the development of a CRISPR-Cas9 host signal reduction method that successfully enriched resulting data for target parasite reads to the extent that clinically-important infections could be detected. Finally, in **Chapter 4,** I systematically reviewed published eukaryotic metabarcoding protocols and undertook the design of a new method, with the intention of capturing all parasite diversity, called VESPA (Vertebrate Eukaryotic endoSymbiont and Parasite Analysis). I optimized the method, compared it to existing protocols, and validated it with a eukaryotic parasite community standard that I created. I then applied the method to clinically relevant samples, compared its performance to the gold standard method of microscopic examination, concluding that the new metabarcoding-based method resulted in higher sensitivity (probability of detection) and taxonomic resolution.

Future directions

Discovery of the novel disease-associated bacterium *Sarcina troglodytae* and initial characterization of epizootic neurologic and gastroenteric syndrome in chimpanzees opens up new avenues for further investigation. Continued research efforts are important for the health of the chimpanzees in Sierra Leone, but also potentially for other species, as evidenced by subsequent outbreaks of similar disease associated with *Sarcina* organisms in animals ([1] and unpublished communications) and by the continued steady stream of published cases in humans [2-5]. Future work should include epidemiological characterization of such outbreaks along with more widespread testing for *Sarcina* bacteria, including genotyping to definitively identify *Sarcina* species. Specifically in the case of *S. troglodytae,* it is critical to find ideal growth conditions for establishing stable, long-term cultures necessary for in-depth lab study, which should include further characterization by antimicrobial sensitivity testing and application to an infection model. Stable cultures would also enable autogenous vaccine development, which has been successful for control of bacterial infections in other veterinary contexts [6, 7].

With regards to new eukaryotic endosymbiont methods, I am most excited about the varied and impactful applications of VESPA and CRISPR-Cas9 enrichment to fascinating, long-standing biological questions. There are fundamental aspects of endosymbiont assemblage formation and function of which our knowledge is limited or nonexistent that can now be addressed with metabarcoding (reviewed in [8, 9]). For example: how stable are gut eukaryotic endosymbiont communities over time [10]? How and when in the life course of a host are gut eukaryotic endosymbiont communities established [11] and do they change over time? Are certain endosymbiont species or combinations of species associated with positive or negative health outcomes [12]? How does treatment with various antiparasitic drugs or immunosuppression impact community structure in the short-term and long-term [13]?

Furthermore, there are diagnostic applications for VESPA that may be impactful for veterinary and human medicine. In clinical cases where rapid assessment of multiple potential eukaryotic pathogens is needed, but where there is limited evidence to guide testing for specific organisms, unbiased VESPA could be used to identify all organisms present [14]. This is similar to the expanding use of microbiome and virome analyses, originally developed for research purposes, in human and animal clinical diagnostics [15] (reviewed in [16]). Likewise, in outbreak investigations where traditional diagnostics have failed, VESPA could provide a broad look at possible disease agents. Coupled with appropriate epidemiological study designs, this information could identify candidate organisms for further study or development of targeted diagnostics. Such approaches are already being developed for food-borne bacterial outbreaks [17] and could be extended to eukaryotic pathogens as well.

There are also potential uses for eukaryotic endosymbiont metabarcoding in healthy populations. For example, it is important to establish baseline data for parasites in domestic animals, production animals, and wildlife populations for comparison in the case of future disease outbreaks , as has been achieved for viruses in some contexts [18]. Metabarcoding could also be used as a surveillance tool to look for circulating parasitic pathogens and help predict both adverse events to host health and the potential for pathogen spillover to other species, including humans [19].

The design of VESPA is purposefully compatible with other, existing metabarcoding platforms. Thus far, cross-kingdom analyses have included viruses, bacteria, archaea, and fungi but have lacked data on protozoa and helminths. With the addition of VESPA for this “missing link,” the full picture of organism communities will be possible, and we can address such questions as: does eukaryotic endosymbiont community composition and diversity covary with bacterial community composition and diversity (reviewed in [20])? Do the establishments of bacterial and eukaryotic communities during host development impact each other [21]? What are the effects of various drug treatments on both bacterial and eukaryotic communities [22]?

Through my thesis work, I have developed new tools which I envision can help move the fields of parasitology and microbial ecology forward and enable new insights into eukaryotic endosymbiont communities in human and animal hosts. In turn, these insights will enable improved diagnostics, targeted therapeutics, and more potent preventive health strategies. I hope that the application of metabarcoding technology will allow for insights akin to those made through bacterial and fungal metabarcoding and resulting discoveries.

References

1. Tuzcu M, Tuzcu N, Akcakavak G, Celik Z. Diagnosis of *Sarcina ventriculi*-derived haemorrhagic abomasitis in lambs by histopathology and real-time PCR. Acta Vet Brno. 2022;91(3):227-33.

2. Tartaglia D, Coccolini F, Mazzoni A, Strambi S, Cicuttin E, Cremonini C, et al. *Sarcina ventriculi* infection: a rare but fearsome event. A Systematic Review of the Literature. Int J Infect Dis. 2022;115:48-61. Epub 2021/11/29. doi: 10.1016/j.ijid.2021.11.027. PubMed PMID: 34838720.

3. de la Fuente Molinero I, MT BR, MP AG, Arbide Del Río N. *Sarcina ventriculi* in gastric biopsies of two patients with an underlying neoplasia. Revista Espanola de Enfermedades Digestivas: Organo Oficial de la Sociedad Espanola de Patologia Digestiva. 2022;115.

4. Vuković AS, Jonjić N, Veršić AB, Kovač D, Radman M. Fatal Outcome of Emphysematous Gastritis due to *Sarcina ventriculi* Infection. Case Rep Gastroenterol. 2021;15:933-8.

5. Prajapati RM, Nandwani SK, Kabrawala MV, Patel NB, Arora PV, Parekh KK. *Sarcina Ventriculi* of Gastrointestinal Tract: A Clinicopathologic Study. Trop Gastroenterol. 2022;42(4):181-8.

6. Corsaut L, Misener M, Canning P, Beauchamp G, Gottschalk M, Segura M. Field Study on the Immunological Response and Protective Effect of a Licensed Autogenous Vaccine to Control Streptococcus suis Infections in Post-Weaned Piglets. Vaccines (Basel). 2020;8(3). Epub 2020/07/18. doi: 10.3390/vaccines8030384. PubMed PMID: 32674276; PubMed Central PMCID: PMCPMC7565864.

7. Kromann S, Olsen RH, Bojesen AM, Jensen HE, Thofner I. Protective Potential of an Autogenous Vaccine in an Aerogenous Model of Escherichia coli Infection in Broiler Breeders. Vaccines (Basel). 2021;9(11). Epub 2021/11/28. doi: 10.3390/vaccines9111233. PubMed PMID: 34835164; PubMed Central PMCID: PMCPMC8624668.

8. Parfrey LW, Walters WA, Knight R. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. Frontiers in Microbiology. 2011;2. doi: 10.3389/fmicb.2011.00153. PubMed PMID: 21808637 PMCID: PMC3135866.

9. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258-70. Epub 2012/03/20. doi: 10.1016/j.cell.2012.01.035. PubMed PMID: 22424233; PubMed Central PMCID: PMCPMC5050011.

10. Scanlan PD, Marchesi JR. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. Isme J. 2008;2(12):1183-93. Epub 2008/08/02. doi: 10.1038/ismej.2008.76. PubMed PMID: 18670396.

11. Valzania L, Martinson VG, Harrison RE, Boyd BM, Coon KL, Brown MR, et al. Both living bacteria and eukaryotes in the mosquito gut promote growth of larvae. PLoS Negl Trop Dis. 2018;12(7):e0006638.

12. Betts EL, Gentekaki E, Tsaousis AD. Exploring Micro-Eukaryotic Diversity in the Gut: Co-occurrence of Blastocystis Subtypes and Other Protists in Zoo Animals. Front Microbiol. 2020;11:288. Epub 2020/03/13. doi: 10.3389/fmicb.2020.00288. PubMed PMID: 32161577; PubMed Central PMCID: PMCPMC7052370.

13. Hamad I, Abou Abdallah R, Ravaux I, Mokhtari S, Tissot-Dupont H, Michelle C, et al. Metabarcoding analysis of eukaryotic microbiota in the gut of HIV-infected patients. PLoS One. 2018;13(1):e0191913. Epub 2018/02/01. doi: 10.1371/journal.pone.0191913. PubMed PMID: 29385188; PubMed Central PMCID: PMCPMC5791994.

14. Brown JR, Bharucha T, Breuer J. Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. J Infect. 2018;76(3):225-40. Epub 2018/01/07. doi: 10.1016/j.jinf.2017.12.014. PubMed PMID: 29305150; PubMed Central PMCID: PMCPMC7112567.

15. Babiker A, Bradley H, Stittleburg V, Key A, Kraft CS, Waggoner J, et al. Metagenomic sequencing to detect respiratory viruses in persons under investigation for COVID-19. medRxiv. 2020. Epub 2020/09/17. doi: 10.1101/2020.09.09.20178764. PubMed PMID: 32935115; PubMed Central PMCID: PMCPMC7491530.

16. Jia X, Hu L, Wu M, Ling Y, Wang W, Lu H, et al. A streamlined clinical metagenomic sequencing protocol for rapid pathogen identification. Sci Rep. 2021;11(1):4405. Epub 2021/02/25. doi: 10.1038/s41598-021-83812-x. PubMed PMID: 33623127; PubMed Central PMCID: PMCPMC7902651.

17. Buytaers FE, Saltykova A, Mattheus W, Verhaegen B, Roosens NHC, Vanneste K, et al. Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a Salmonella food-borne outbreak. Microb Genom. 2021;7(4). Epub 2021/04/08. doi: 10.1099/mgen.0.000547. PubMed PMID: 33826490; PubMed Central PMCID: PMCPMC8208685.

18. Bergner LM, Orton RJ, da Silva Filipe A, Shaw AE, Becker DJ, Tello C, et al. Using noninvasive metagenomics to characterize viral communities from wildlife. Mol Ecol Resour. 2019;19(1):128-43. Epub 2018/09/22. doi: 10.1111/1755-0998.12946. PubMed PMID: 30240114; PubMed Central PMCID: PMCPMC6378809.

19. Titcomb GC, Jerde CL, Young HS. High-Throughput Sequencing for Understanding the Ecology of Emerging Infectious Diseases at the Wildlife-Human Interface. Front Ecol Evol. 2019;7. doi: 10.3389/fevo.2019.00126.

20. Deng L, Wojciech L, Gascoigne NRJ, Peng G, Tan KSW. New insights into the interactions between Blastocystis, the gut microbiota, and host immunity. PLoS Pathog. 2021;17(2):e1009253. Epub 2021/02/26. doi: 10.1371/journal.ppat.1009253. PubMed PMID: 33630979; PubMed Central PMCID: PMCPMC7906322.

21. Wampach L, Heintz-Buschart A, Hogan A, Muller EEL, Narayanasamy S, Laczny CC, et al. Colonization and Succession within the Human Gut Microbiome by Archaea,

Bacteria, and Microeukaryotes during the First Year of Life. Front Microbiol. 2017;8:738. Epub 2017/05/18. doi: 10.3389/fmicb.2017.00738. PubMed PMID: 28512451; PubMed Central PMCID: PMCPMC5411419.

22. Bär A-K, Phukan N, Pinheiro J, Simoes-Barbosa A. The interplay of host microbiota and parasitic protozoans at mucosal interfaces: implications for the outcomes of infections and diseases. PLoS Negl Trop Dis. 2015;9(12):e0004176.