



# VESPA: an optimized protocol for accurate metabarcoding-based characterization of vertebrate eukaryotic endosymbiont and parasite assemblages

Received: 4 April 2023

Accepted: 15 December 2023

Published online: 09 January 2024

 Check for updates

Leah A. Owens<sup>1</sup> , Sagan Friant<sup>1,2,3</sup>, Bruno Martorelli Di Genova<sup>4,5</sup>,  
Laura J. Knoll<sup>4</sup>, Monica Contreras<sup>6</sup>, Oscar Noya-Alarcon<sup>7</sup>,  
Maria G. Dominguez-Bello<sup>8,9,10,11</sup> & Tony L. Goldberg<sup>1</sup> 

Protocols for characterizing taxonomic assemblages by deep sequencing of short DNA barcode regions (metabarcoding) have revolutionized our understanding of microbial communities and are standardized for bacteria, archaea, and fungi. Unfortunately, comparable methods for host-associated eukaryotes have lagged due to technical challenges. Despite 54 published studies, issues remain with primer complementarity, off-target amplification, and lack of external validation. Here, we present VESPA (Vertebrate Eukaryotic endo-Symbiont and Parasite Analysis) primers and optimized metabarcoding protocol for host-associated eukaryotic community analysis. Using in silico prediction, panel PCR, engineered mock community standards, and clinical samples, we demonstrate VESPA to be more effective at resolving host-associated eukaryotic assemblages than previously published methods and to minimize off-target amplification. When applied to human and non-human primate samples, VESPA enables 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.

Microbiomes are multikingdom assemblages of microorganisms and their entire “theater of activity” including signaling molecules and metabolites<sup>1</sup>. Such communities have emergent properties arising from cross-species and cross-kingdom interactions<sup>2</sup>. One of the most

salient examples is the human gut, wherein bacterial community dynamics have direct effects on health<sup>3</sup> and can be manipulated to improve disease outcomes in clinical settings<sup>4</sup>. Evidence is mounting that assemblages of host-associated eukaryotes also form

<sup>1</sup>Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, USA. <sup>2</sup>Department of Anthropology, The Pennsylvania State University, University Park, PA, USA. <sup>3</sup>Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA. <sup>4</sup>Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI, USA. <sup>5</sup>Department of Microbiology and Molecular Genetics, Larner College of Medicine, The University of Vermont, Burlington, VT, USA. <sup>6</sup>Center for Biophysics and Biochemistry, Venezuelan Institute of Scientific Research (IVIC), Caracas, Venezuela. <sup>7</sup>Centro Amazónico de Investigación y Control de Enfermedades Tropicales-CAICET, Puerto Ayacucho, Amazonas, Venezuela. <sup>8</sup>Department of Biochemistry and Microbiology, Rutgers University–New Brunswick, New Brunswick, NJ, USA. <sup>9</sup>Department of Anthropology, Rutgers University, New Brunswick, NJ, USA. <sup>10</sup>Institute for Food, Nutrition and Health, Rutgers University, New Brunswick, NJ, USA. <sup>11</sup>Canadian Institute for Advanced Research (CIFAR), Toronto, ON, Canada. ✉e-mail: [leah.owens@wisc.edu](mailto:leah.owens@wisc.edu); [tony.goldberg@wisc.edu](mailto:tony.goldberg@wisc.edu)

communities with important consequences for host health<sup>5</sup>, although they are far less studied compared to their bacterial, archaeal, and fungal counterparts<sup>6</sup>. Even terminology to describe host-associated eukaryotes is lacking. “Eukaryotic microbiome/microbiota”<sup>7</sup> does not include host-associated macro-organisms such as helminths, “nemabiome”<sup>8</sup> is limited to nematodes, and “parasites”<sup>9</sup> excludes commensal/beneficial organisms and includes ectoparasites. Herein we use the term “eukaryotic endosymbionts” to refer to both host-associated 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<sup>10</sup> and the fact that established methods already exist for assessing the “mycobiome”<sup>11</sup>. However, we include microsporidia, because their life cycles are considered more similar to protozoa than to fungi<sup>12</sup>.

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<sup>13</sup>. Combined with subsequent advances in staining and enrichment techniques, microscopy is still a gold standard method<sup>14</sup>, although it requires specialized training<sup>15</sup> and has inherent resolution limits (i.e., some species cannot be distinguished solely based on morphology, a phenomenon known as “cryptic species complexes”<sup>16</sup>). For example, the genus *Entamoeba* contains pathogenic *E. histolytica* and benign *E. dispar* which appear identical under the microscope<sup>17</sup>. 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<sup>18</sup>. Although 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)<sup>19</sup>. For bacteria, the 16S ribosomal RNA (16S rRNA, or just 16S) locus<sup>20</sup> and for fungi, the internal transcribed spacer (ITS) locus<sup>21</sup> are proven targets for metabarcoding. By contrast, universal targets and protocols for metabarcoding of eukaryotic endosymbionts are not standardized<sup>6</sup>. For example, some published methods utilize PCR primer sets originally designed for free-living eukaryotic microbes<sup>22–25</sup>, some target metazoans only<sup>26,27</sup>, while others focus exclusively on helminths<sup>8,28–30</sup> or gut-associated organisms<sup>31–33</sup>. There is also a conspicuous absence of published comparisons to gold standard methods such as microscopy<sup>34</sup>. 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<sup>35</sup>. Unfortunately, no such standard exists for eukaryotes other than fungi.

Here we present and validate the VESPA (Vertebrate Eukaryotic endoSymbiont and Parasite Analysis) primers and optimized protocol for eukaryotic endosymbiont metabarcoding that resolves the issues described above. We compare VESPA to published methods in silico and using community standards comprised of cloned DNA from eukaryotic endosymbiont lineages across the Tree of Life. We then quantify off-target signal abundance and finally compare our protocol 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 protocol of our own design in a progressive series of experiments. We begin with in silico PCR, proceed to amplification of single parasite DNA templates, conduct metabarcoding using two engineered mock community standards, and assess off-target signals in fecal samples. 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 protocol design

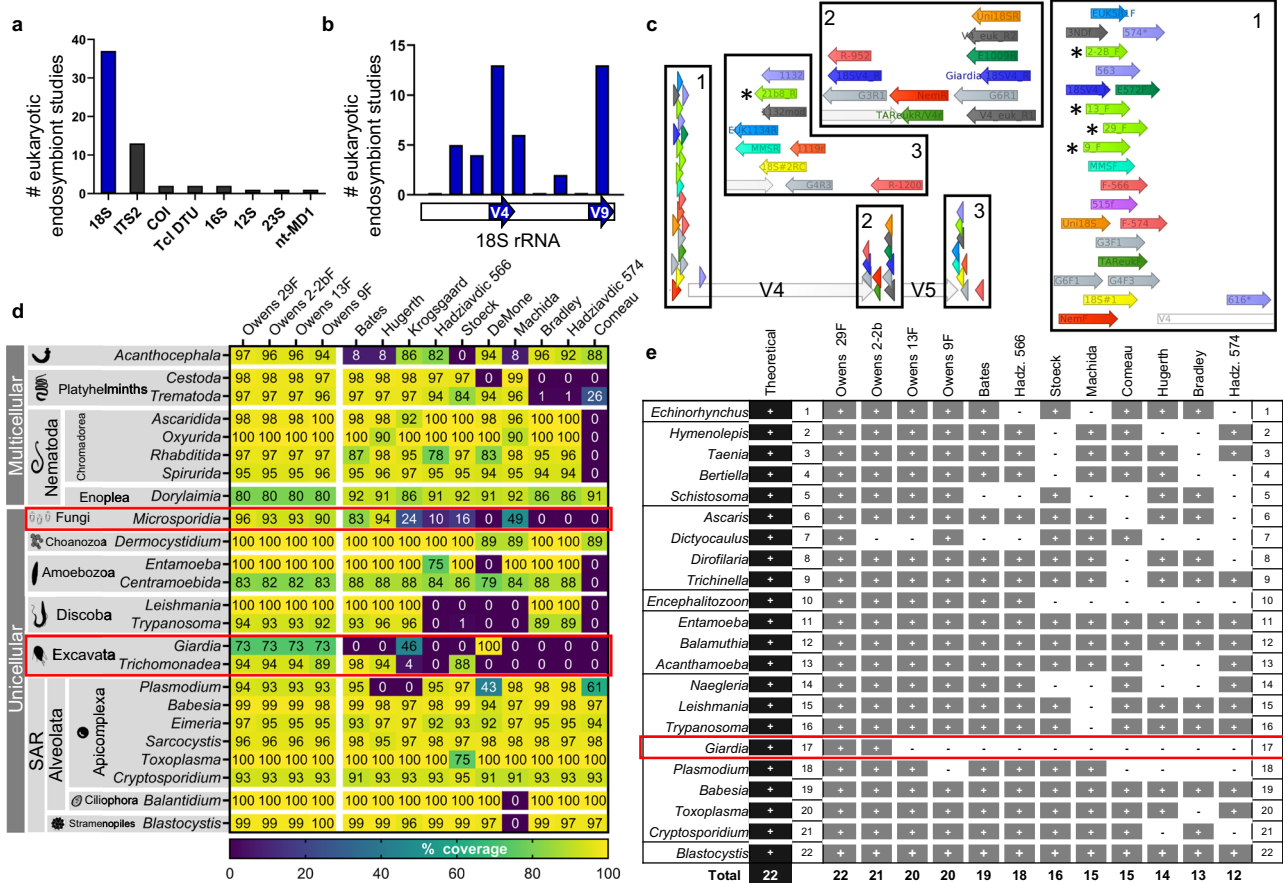
In a literature review consisting of 54 papers that used amplicon sequencing (metabarcoding) to characterize eukaryotic assemblages in vertebrate hosts (Supplementary Data 1), we identified eight marker genes, including nt-MD1 ( $n=1$ ), 12S ( $n=1$ ), 28S ( $n=1$ ), mitochondrial 16S ( $n=2$ ), mini-exon TcI DTU ( $n=2$ ), COI ( $n=2$ ), ITS-2 ( $n=13$ ), and 18S ( $n=37$ ; Fig. 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<sup>36</sup>, 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 ( $n=13$ ; Fig. 1b). The 18S V4 region has the highest entropy within the size limits of MiSeq v2 chemistry<sup>37</sup> 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 18S V4 primers designed to target all vertebrate eukaryotic endosymbionts, consisting of 4 candidate forward primers and one reverse primer (see Methods section for details on primer design, Supplementary Table 1 for primer sequences and Fig. 1c for a map of primer binding sites).

### Testing metabarcoding methods for taxonomic coverage and resolution using in silico PCR

Testing all 22 published 18S V4 primer sets in silico (condensed for coverage analysis by combining primer sets in 2 cases for 19 total Primer IDs) yielded an average eukaryotic endosymbiont coverage of 64.9% (Table 1, bolded columns). No primer set recognized both *Plasmodium* and *Giardia*, and 7 sets 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, 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 prokaryotic coverage were not analyzed further as prokaryotes constitute the vast majority of genetic material in gut flora and fecal samples<sup>38</sup>.

In silico PCR including our four primer sets alongside the remaining 18 published 18S V4 sets (condensed in 2 cases, total 19 Primer IDs) yielded coverage data spanning a wide range (5.7–98.0%; Table 2). Across target groups (normalizing by eligible accessions), our 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, penultimate column). Overall coverage of fungal groups was high with all primer sets (45.1–94.1%, mean 86.2%; Table 2, final column), as was expected based on 18S sequence similarity between fungi and eukaryotic endosymbionts<sup>39</sup>.



**Fig. 1 | VESPA development and evaluation.** **a** Histogram of marker genes identified in a literature review of 54 host-associated eukaryotic endosymbiont studies (see Supplementary Data 1). **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 V4–V5 (open arrows) of eukaryotic 18S SSU rRNA gene. Newly designed (asterisks) and published metabarcoding primer sets are shown as colored arrows and boxed areas 1–3 are expanded as insets. See Supplementary Table 1 for full primer names and sequences. **d** Heat map of published and new 18S

V4 primer set coverage across clades exclusively containing eukaryotic endosymbionts 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 eukaryotic endosymbiont 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.

In silico coverage analysis using finer-resolution groups (Fig. 1d) showed that our designed 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; Fig. 1d, red boxes).

We tested the taxonomic resolution of the 18S V4 region amplified by our primers by evaluating the amplicon sequences of known eukaryotic vertebrate endosymbionts generated by in silico PCR ( $n = 3769$ ). Overall, 98.3% of sequences ( $n = 3704$ ) could be resolved to the species level, 1.4% ( $n = 52$ ) to the genus level, and 0.3% ( $n = 13$ ) to the family level (Table 3). Normalized by total sequences, Platyhelminthes had the greatest number of sequences not resolved to the species or genus level, but with only 6 of 531 (1.13%) sequences assigned to more than one genus (Supplementary Table 2).

### Testing metabarcoding methods for on-target amplification using purified DNA

In PCR amplification of genomic DNA (gDNA) from 22 individual eukaryotic endosymbiont organisms (Supplementary Table 3), 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 (Fig. 1e). Furthermore, two of the new sets were the only primers to successfully amplify 18S V4 from *Giardia* gDNA (Owens 29 F and Owens 2-2bF), as expected based on in silico data (Fig. 1e, red box).

### Testing metabarcoding methods for amplification bias using a community standard

Community standards are not available for eukaryotic endosymbionts, so we collected host-associated protozoa ( $n = 10$ ), helminths ( $n = 5$ ), and a microsporidian ( $n = 1$ ) (Supplementary Table 3) from various sources (e.g., specimen repositories, veterinary post-mortem examinations). We then isolated 18S genes from these samples and combined them at different relative abundances to create two community standards (Fig. 2a), one consisting of equal 18S gene copies from each organism, which we named Equimolar EukMix, and one with unequal 18S gene copies following a logarithmic function (maximum 400-fold abundance difference), which we named Log EukMix (Supplementary Table 4). Metabarcoding Equimolar EukMix and Log EukMix with previously published and newly designed primers allowed us to directly compare empirical read abundances for each organism to

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

Primer ID	n = Primer sets	Off-target groups		Eukaryotic endosymbiont groups		Specific examples	
		20,197 Archaea	381,535 Bacteria	4229 Helminths	15,265 Protozoa	198 <i>Plasmodium</i>	23 <i>Giardia</i>
Bates <sup>106</sup>	515 f/1119r	0	0	<b>80.4</b>	<b>95.9</b>	94.8	0
Bower <sup>a 107</sup>	18SEUK581F/18SEUK1134R	46.2 <sup>a</sup>	8.2 <sup>a</sup>	<b>0.4</b>	<b>82.4</b>	0	72.7
Bradley <sup>37</sup>	TAReuk454FWD1/V4r	0	0	<b>48.9</b>	<b>67.1</b>	97.9	0
C-S <sup>108</sup> /Brate 2 <sup>109</sup>	3NDf/V4_euk_R2	0	0	<b>50.8</b>	<b>22.8</b>	0	0
C-S <sup>108</sup> /Brate 1 <sup>109</sup>	3NDf/V4_euk_R1	0	0	<b>5.8</b>	<b>21.1</b>	0	0
C-S <sup>108</sup> /Geisen <sup>110</sup>	3NDf/1132mod	0.3	0	<b>80.7</b>	<b>94.2</b>	0	0
Comeau <sup>111</sup>	E572F/E1009R	0	0	<b>65.3</b>	<b>44.5</b>	0	0
DeMone <sup>b112</sup>	18SV4_F/_R/Giardia_R	0	0	<b>86.4</b>	<b>62.3</b>	0	100
Hadziavdic 566 <sup>73</sup>	F-566/R-1200	0	0	<b>76.4</b>	<b>81</b>	99.6	0
Hadziavdic 574 <sup>73</sup>	F-574/R-952	0	0	<b>48.3</b>	<b>62.9</b>	61.3	0
Hugerth 574 <sup>a113</sup>	574/1132	12.5 <sup>a</sup>	0	<b>80</b>	<b>94.2</b>	0	0
Hugerth 616 <sup>113</sup>	616/1132	3.3	0.2	<b>93.1</b>	<b>75.8</b>	0	45.5
Hugerth 563 <sup>a113</sup>	563/1132	47.9 <sup>a</sup>	72 <sup>a</sup>	<b>96.1</b>	<b>96.4</b>	0	100
Krogsgaard <sup>b32</sup>	G3F1/R1/G4F3/R3/G6F1/R1	0	0	<b>78.5</b>	<b>67</b>	94.8	0
Machida <sup>114</sup>	18S#1/18S#2RC	0	0	<b>78.1</b>	<b>45.2</b>	97.9	0
Sikder <sup>a115</sup>	MMSF/MMSR	17.5 <sup>a</sup>	0	<b>79.3</b>	<b>42.7</b>	0	0
Stoeck <sup>116</sup>	TAReuk454F1/R3	0	0	<b>49.1</b>	<b>78.4</b>	97.9	0
Wood <sup>117</sup>	Nem18SlongF/Nem18SlongR	0	0	<b>32.2</b>	<b>25.2</b>	2.6	0
Zhan <sup>118</sup>	Uni18S/R	0	0	<b>72.8</b>	<b>64</b>	0	0

Bold font denotes target organisms. See Supplementary Table 1 for full primer names and sequences.

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.

<sup>a</sup>Removed from further analysis due to high prokaryotic complementarity.

<sup>b</sup>Multiple primer sets were combined for analysis.

their predicted abundances (Fig. 2b, top panel) over a wide range of 18S gene abundances. Analysis of Equimolar EukMix and Log EukMix metabarcoding reads demonstrated that the mean fold difference from the theoretical input (a reflection of PCR bias<sup>37</sup>) was significantly lower with designed primer set Owens 29F (Fig. 2b, bottom panel) compared to all published primer sets (see Supplementary Table 5 for exact *P*-values). Similarly, overall abundance distributions resulting from primer set Owens 29F were strongly positively correlated (a reflection of accurate community reconstruction<sup>40,41</sup>) to that of both mock community inputs (Equimolar EukMix:  $r = 0.938$ , Log EukMix:  $r = 0.9554$ ; Pearson correlation coefficient) and more closely approximated the actual community than any other primer set tested (Fig. 2c). We therefore chose 29F/21b8R as the primer set for our finalized VESPA metabarcoding protocol (see Supplementary Note 1–VESPA protocol).

### Testing VESPA for off-target amplification of prokaryotic, fungal, and host sequences

To test the performance and off-target amplification abundance of VESPA primers when applied to clinically-relevant samples, we used DNA extracted from whole human feces ( $n = 40$ ) as input for metabarcoding, which yielded a mean 44,311 reads per sample after quality-filtering (range: 23,093–89,510; Fig. 3a). Reads originating from bacteria and archaea accounted for a mean 3.8% (range: 0–28.5%), host accounted for a mean 6.4% (range: 0–32.1%), and fungi accounted for a mean 3.1% (range: 0–29.7%; Fig. 3b). Reads originating from on-target organisms (Fig. 3a) accounted for an overall mean of 86.7% (range: 55.1–99.4%; Fig. 3b).

### VESPA compared to microscopy in human samples

VESPA analysis of 12 human clinical samples yielded high-quality data (Supplementary Table 6) including low proportions of off-target

prokaryotic reads (Fig. 4a) and host reads (host read mean = 2.97% per sample, range: 0.11–17.4%) and correspondingly high proportions of endosymbiont reads (Fig. 4b, c).

VESPA successfully identified all three helminth and seven protozoan taxa identified with microscopy (Fig. 4d) and found these taxa in more individuals than did microscopy, with 61.4% of positive samples identified solely by VESPA (Fig. 4e). 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,  $P = 0.0005$ ; Fig. 4f, 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,  $P = 0.002$ ; Fig. 4f, right panel).

### VESPA compared to microscopy in non-human primate samples

VESPA analysis of 40 non-human primate clinical samples yielded high-quality sequencing reads (Supplementary Table 6) with low proportions of off-target prokaryotic reads (Fig. 5a) and host sequence reads (host read mean = 3.2% per sample, range: 0%–18.49%) and correspondingly high proportions of endosymbiont reads (Fig. 5b, c).



**Table 2 | In silico taxonomic coverage of host-associated helminths/protozoa and fungi for published and newly designed 18S V4 primer sets**

Primer ID	Mean	n = 3097		n = 2913		Rank		Rank sum	n = 9373
		Helminths	Protozoa	Helminths	Protozoa	Helminths	Protozoa		
Owens 29F	<b>96.8%</b>	95.5%	98.0%	1	1	<b>2</b>		91.1%	
Owens 2-2b	<b>96.4%</b>	94.9%	97.9%	2	2	<b>4</b>		92.2%	
Owens 13F	<b>96.4%</b>	94.9%	97.9%	2	2	<b>4</b>		92.4%	
Owens 9F	<b>95.2%</b>	94.4%	96.0%	4	4	<b>8</b>		92.4%	
Bates	<b>88.2%</b>	80.4%	95.9%	8	5	<b>13</b>		93.7%	
Hugerth	<b>84.5%</b>	93.1%	75.8%	5	8	<b>13</b>		94.1%	
Krogsgaard <sup>a</sup>	<b>81.0%</b>	93.0%	69.0%	6	9	<b>15</b>		91.8%	
Hadziavdic 566	<b>78.7%</b>	76.4%	81.0%	10	6	<b>16</b>		91.4%	
DeMone <sup>a</sup>	<b>75.3%</b>	86.4%	64.1%	7	11	<b>18</b>		81.8%	
Stoeck	<b>63.8%</b>	49.1%	78.4%	13	7	<b>20</b>		79.4%	
Machida	<b>61.7%</b>	78.1%	45.2%	9	14	<b>23</b>		92.5%	
Bradley	<b>58.0%</b>	48.9%	67.1%	14	10	<b>24</b>		69.5%	
Hadziavdic 574	<b>55.6%</b>	48.3%	62.9%	15	12	<b>27</b>		78.9%	
Comeau	<b>54.9%</b>	65.3%	44.5%	11	15	<b>26</b>		91.4%	
C-S/Geisen <sup>b</sup>	<b>47.5%</b>	40.7%	54.2%	16	13	<b>29</b>		93.0%	
C-S/Brate 2 <sup>b</sup>	<b>36.8%</b>	50.8%	22.8%	12	17	<b>29</b>		90.0%	
Wood <sup>b</sup>	<b>28.7%</b>	32.2%	25.2%	17	16	<b>33</b>		45.1%	
Zhan <sup>b</sup>	<b>14.0%</b>	5.7%	22.3%	19	18	<b>37</b>		83.6%	
C-S/Brate 1 <sup>b</sup>	<b>13.5%</b>	5.8%	21.1%	18	19	<b>37</b>		93.8%	

Bold font indicates overall metrics (mean and rank sum). See Supplementary Table 1 for full primer names and sequences.

The first four rows represent 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.

<sup>a</sup>Multiple primer sets combined for analysis.

<sup>b</sup><50% overall mean target complementarity.

**Table 3 | In silico 18S V4 taxonomic resolution of host-associated protozoa and helminths**

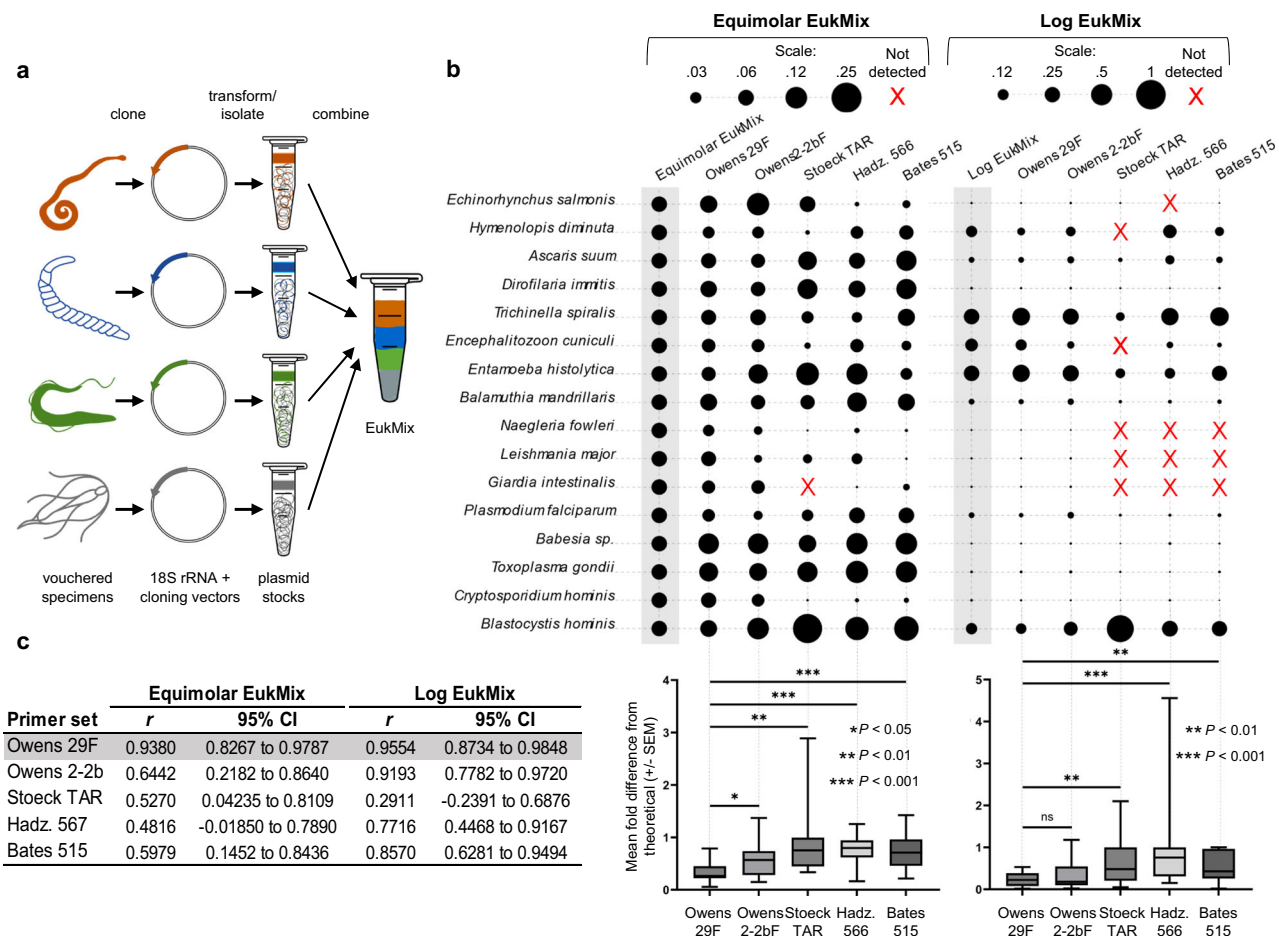
	Total unique sequences	Count			Percent		
		1 species	>1 species	>1 genus	1 species	>1 species	>1 genus
Blastocystis	141	139	2	0	0.986	0.014	0.000
Ciliophora	772	770	2	0	0.997	0.003	0.000
Apicomplexa	1476	1458	17	1	0.988	0.012	0.001
Amoebozoa	317	316	1	0	0.997	0.003	0.000
Acanthocephala	72	69	3	0	0.958	0.042	0.000
Platyhelminthes	531	517	6	8	0.974	0.011	0.015
Nematoda	460	435	21	4	0.946	0.046	0.009

Third and sixth columns represent species-level resolution.

See Supplementary Table 2 for full list of unresolved species.

VESPA successfully identified all eight helminth and six protozoan taxa identified with microscopy (Fig. 5d) and found these taxa in more individuals than did microscopy, with 47.08% of positive samples identified by VESPA only (Fig. 5e). 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; Fig. 5d, e). *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, constituting resolution of a cryptic species complex. 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,  $P = 0.25$ ; Fig. 5f, 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 Fig. 5f, right panel).



**Fig. 2 | Testing metabarcoding methods for amplification bias using community standards. a** Schematic overview of EukMix creation via 18S isolation and cloning. **b** Equimolar EukMix (left panel) and Log EukMix (right panel) community standard metabarcoding across primer sets (columns, labels at top and bottom) as compared to theoretical input (leftmost columns, gray boxes) shown as % abundance of reads per organism corresponding to bubble area. Any organism not detected (0 reads after quality filtering) is shown as a red X. See Supplementary Table 3 for parasite sources and strains. Bottom half of panels shows box plots of mean fold difference from theoretical input of three replicates  $\pm$  standard error of

the mean (SEM) in which the central line is the median, box limits indicate first to third quartile range, and whiskers display minimum and maximum values. *P* values are derived from two-tailed Wilcoxon matched-pairs signed rank tests. ns, not significant. See Supplementary Table 5 for exact *P*-values. **c** Pearson correlation coefficients for the relationship between theoretical and observed read abundances for Equimolar EukMix (top) and Log EukMix (bottom). Primer set Owens 29F recovered the underlying communities most accurately (shading). Source data are provided as a Source Data file.

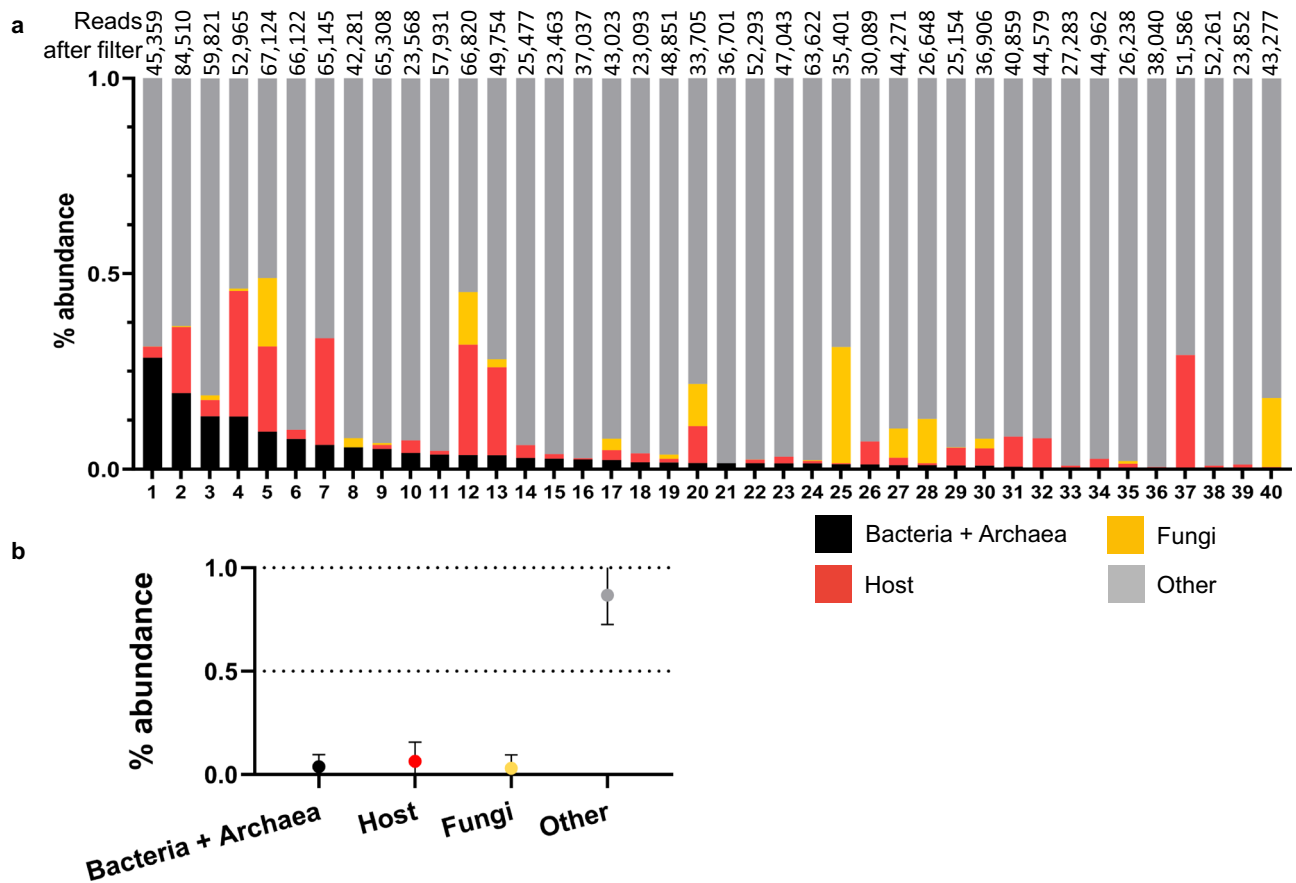
## Discussion

To identify a single protocol 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 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 the designed primer set 29F recognizes the greatest range of eukaryotic endosymbionts of interest with the least PCR bias of any method tested. When applied to DNA extracted from whole human fecal samples, metabarcoding using the 29F primer set resulted in low abundances of off-target prokaryotic, fungal, and host reads with the majority of reads assigned to eukaryotic endosymbionts of interest. We name our new primers and optimized protocol VESPA (V<sub>er</sub>tebrate E<sub>u</sub>karyotic endoSymbiont and P<sub>ar</sub>asite 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<sup>32,33</sup>. However, this approach adds reagent costs and presents technical challenges related to sequencing and bioinformatics<sup>42,43</sup>. Our single primer set approach should therefore reduce barriers to entry for adopting our new protocol. 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 two mock community standards, Equimolar EukMix and Log EukMix, as input for metabarcoding. In both cases, results from VESPA reflected the underlying composition of the community standard more accurately than did results from other assays. The two EukMix community standards should be useful for quality control in laboratories choosing to adopt our protocol, and for standardization and validation, much as community standards containing bacteria and fungi have enabled standardization of microbiome protocols<sup>35,44</sup>. 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<sup>34,45</sup>.



**Fig. 3 | Testing VESPA for off-target amplification with fecal samples. a** VESPA metabarcoding data shown as percent relative abundance for each organism category. Off-target reads are Bacteria + Archaea, Host, and Fungi. On-target reads

are Other. Read numbers after quality filter ( $Q = 30$ ) for each sample are above the bars. **b** Relative abundance of each organism category from (a) ( $n = 40$  independent samples) displayed as mean  $\pm$  SEM. Source data are provided as a Source Data file.

Copy number corrections have been applied in studies of other systems<sup>46,47</sup>, and such corrections could prove useful for investigations where quantifying organism abundance or biomass are the desired outputs.

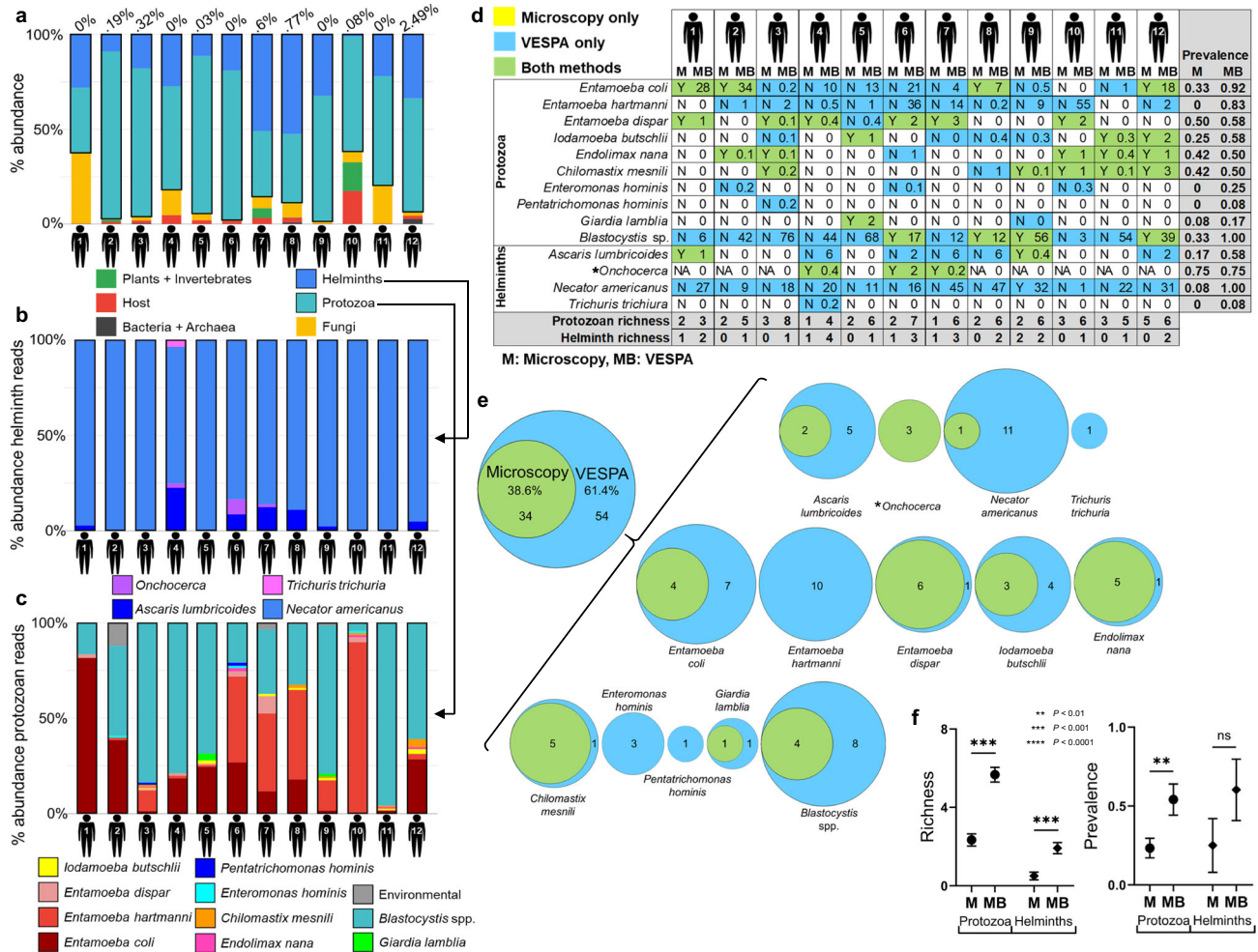
We also note that all marker genes have limits of taxonomic resolution<sup>48</sup>. Use of 18S for metabarcoding affords lower resolution than other, less evolutionarily-conserved marker genes such as mitochondrial cytochrome oxidase 1 (COI)<sup>49,50</sup>, although several commonly cited studies only examined hypervariable regions V1 – V2 or V9<sup>51,52</sup>. Although 18S V4 underperforms compared to COI in some systems<sup>53,54</sup>, recent studies of marine invertebrates have demonstrated that 18S V4 can have comparable species-level resolution to COI<sup>55–60</sup> and may provide more reliable estimates of biomass<sup>61</sup>. Indeed, we found that 98.3% of sequences could be resolved to the species level using our 18S V4 primers. This analysis does not, however, take into account the stochastic effects of metabarcoding in complex samples or the fact that species-level resolution of every eukaryotic endosymbiont group is not possible using 18S<sup>57</sup>. There is an inherent tradeoff between coverage and resolution<sup>62</sup> in marker gene selection that in fact underlies our choice of the 18S gene<sup>63</sup>, with the understanding that some identifications may only be genus level.

When VESPA was applied to human fecal samples to assess on-target vs off-target amplification, a mean 86.7% of resulting reads were identified as on-target reads of interest (Fig. 2d). Overall amplification of fungi was low (mean 3.1%; Fig. 2d), despite high primer cross-complementarity with fungal 18S sequences (Table 2). We suspect this finding to have resulted from the low relative amount of fungal organisms in human feces compared to bacteria and archaea<sup>64</sup>. Fecal

sample composition varies with factors such as host species<sup>65</sup>, diet<sup>66</sup>, age<sup>67</sup>, and immune status<sup>68</sup>, thus we predict that fungal read abundance may be higher in some individuals and some hosts, which may explain the higher percentages of fungal reads in non-human primate samples (Fig. 5) than in human samples.

Compared to microscopic examination, VESPA detected protozoa, microsporidia, and helminths in more individuals, identified additional organisms, differentiated the morphologically identical organisms *Entamoeba histolytica* from *E. dispar*, 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<sup>69</sup>. 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<sup>70</sup>. In this light, we caution that our protocol 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<sup>71,72</sup>.



**Fig. 4 | VESPA compared to microscopy in human clinical samples.** 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 protozoan reads only (archaea, bacteria, host, plants, invertebrates, and fungi removed in (b, c)). a Numbers above bars are the total percentage of prokaryotic (bacterial + archaeal) reads. d 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 two rows, shaded cells) is defined as the total number of species detected by the specified method. Prevalence (final two 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 ( $n = 12$  independent samples) for microscopy (M) and VESPA metabarcoding (MB) findings. Data are shown as mean  $\pm$  SEM. Protozoa richness:  $P = 0.0005$ , Helminths richness:  $P = 0.001$ , Protozoa prevalence:  $P = 0.002$ , Helminths prevalence:  $P = 0.25$ .  $P$  values are derived from Wilcoxon matched-pairs signed rank tests, 2-tailed. ns, not significant. Source data are provided as a Source Data file.

Our contribution with this work is a publicly available protocol for metabarcoding eukaryotic endosymbiont communities that outperforms published methods by all measures 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<sup>5</sup>. VESPA can provide this missing eukaryotic component and thereby enable cross-kingdom characterization of microbial ecosystem

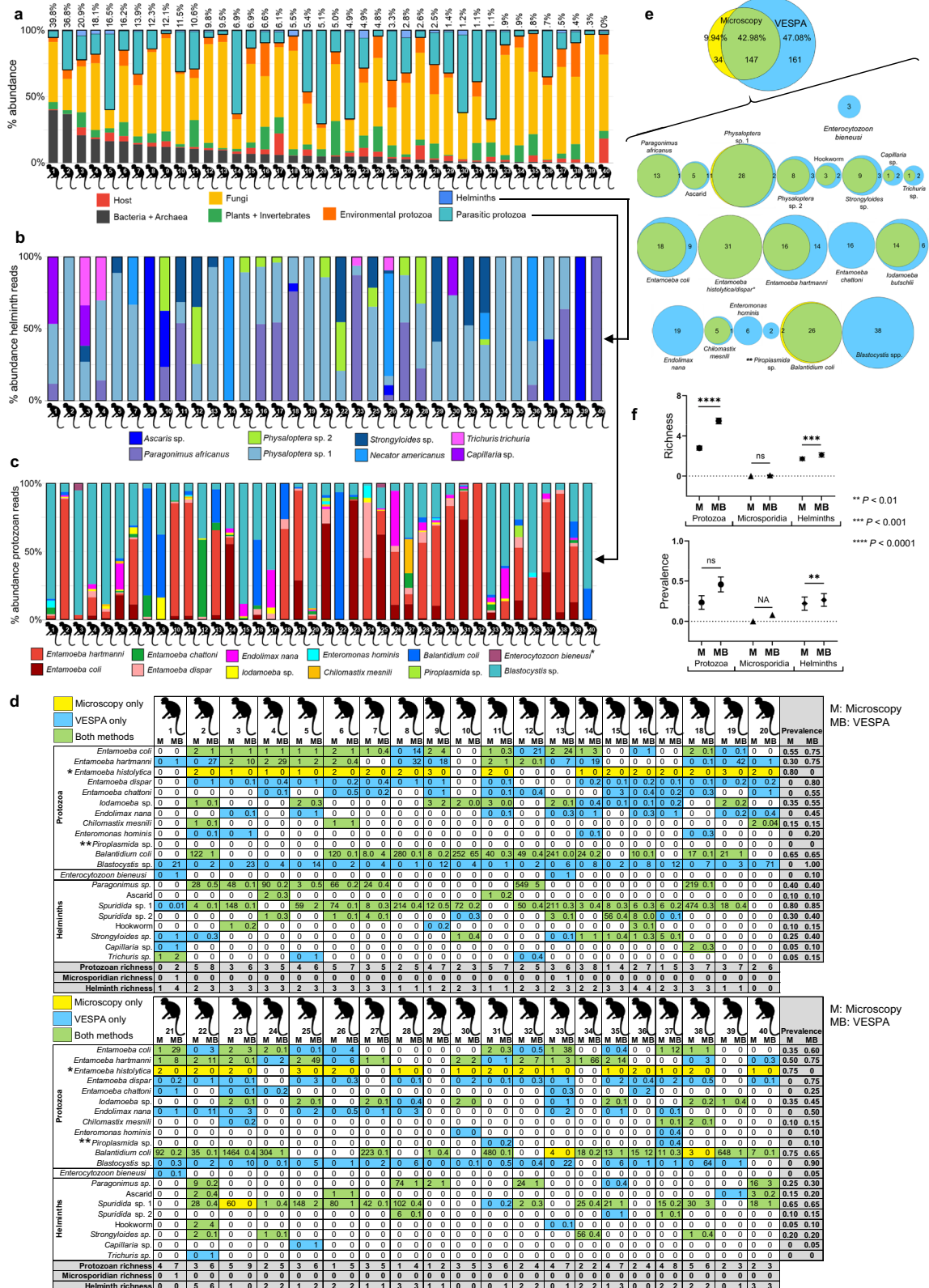
structure and function, opening new avenues for basic and applied research.

## Methods

### Ethical statement

This research complies with all relevant ethical regulations. Clinical samples used in this study were excess material from concluded research and no new samples were collected for this study. Human fecal samples had been collected with appropriate IRB approval (Venezuelan Institute of Scientific Research IRB #DIR-0609/1542/2015 and The University of Wisconsin Madison IRB #2013-1463), including written informed consent in all cases. Non-human primate fecal samples had been collected with appropriate IACUC approval (The University of Wisconsin-Madison's IACUC protocol #VI490). All samples had been completely de-identified prior to use and no demographic or identifying information such as age or sex/gender were provided.





### Methods review and new protocol 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,” “18S,” “short sub unit,” “short subunit,” “short sub-unit,” and “SSU” 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 (Supplementary Data 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<sup>36,49</sup>, 18S has islands of conserved sequence interspersed with

**Fig. 5 | VESPA compared to microscopy in non-human primate clinical samples.** 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, 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 morphologic 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 above 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 ( $n = 40$  independent samples) for microscopy (M) and VESPA (MB) findings. Data are shown as mean  $\pm$  SEM. Protozoa richness:  $P < 0.0001$ , Microsporidian richness:  $P = 0.25$ , Helminths richness:  $P = 0.0009$ , Protozoa prevalence:  $P = 0.0586$ , Helminths prevalence:  $P = 0.0078$ .  $P$  values are derived from Wilcoxon matched-pairs signed rank tests, 2-tailed. ns, not significant. NA, not applicable (single data point only). Source data are provided as a Source Data file.

areas of high entropy (hypervariable regions), allowing broad priming for coverage and diverse amplicons for resolution<sup>73</sup>, and database coverage for 18S is higher than for other loci<sup>74</sup>. Of the 9 hypervariable 18S regions, V4 has the highest taxonomic resolution<sup>37</sup>, so we focused on this region and identified 22 sets of published V4 primers (Supplementary 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<sup>75</sup> or the SILVA 138.1 Small Subunit rRNA Non-Redundant Reference Database ( $n = 510,508$  total accessions;<sup>74,76</sup> 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"<sup>77</sup>. To ensure broad coverage of commensals, zoonoses, and novel organisms we added non-pathogenic protozoans of humans<sup>78</sup>, parasites/commensals of great apes<sup>79</sup>, and parasites of veterinary importance<sup>80</sup>. We then used MUSCLE<sup>81</sup> implemented in MEGA 11<sup>82</sup> 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<sup>83</sup> and the ecoPrimers function in OBIttools<sup>84</sup>, 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<sup>85</sup>. 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<sup>37</sup> 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<sup>86</sup>.

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<sup>87</sup>. 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 (Supplementary Table 1) for further testing.

### Testing metabarcoding methods for taxonomic coverage and resolution 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 (Supplementary Data 1, Supplementary 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 ( $n = 2170$ ), and Platyhelminthes ( $n = 1993$ ) and protozoa accessions included Amoebozoa ( $n = 1148$ ), Discoba ( $n = 1032$ ), Excavata ( $n = 389$ ), Alveolata ( $n = 9140$ ), and Stramenopiles ( $n = 3556$ ). 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 ( $n = 4$ ; Supplementary 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 Fig. 1c for primer map). We compared taxonomic coverage for primer sets using the TestPrime tool<sup>85</sup> and SILVA Ref NR<sup>74,76</sup> allowing for a single mismatch with a mismatch-free two base pair 3' window. Primers with  $\leq 50\%$  overall mean coverage of target groups and methods that required more than a single primer set were not considered further.

Taxonomic resolution of the 18S V4 region amplified by our primer sets was assessed by running the TaxMan server<sup>88</sup> with the Owens 29F primer set as input, no sequence region targeting, a 5% mismatch allowance, a 3 bp 3' mismatch window, and all other parameters set to default values. The database used was SILVA SSU NR rRNA v.128<sup>74</sup>, and data were exported in non-redundant FASTA format, in which all headers contain full taxonomic identifiers and redundant sequences

(100% sequence identity) are concatenated. Sequences corresponding to vertebrate eukaryotic endosymbionts belonging to *Blastocystis*, Ciliophora, Apicomplexa, Amoebozoa, Acanthocephala, Platyhelminthes, and Nematoda were retrieved and binned at the species level (single species in taxonomic header ID), the genus level (more than 1 species of the same genus in taxonomic header ID), or the family level (more than one genus in the taxonomic header ID). All unique sequences with more than one species ID (i.e., sequences that could not be resolved to the species level) are shown in Supplementary Table 2.

### Testing metabarcoding methods for on-target amplification using purified DNA

We assessed amplification success of the remaining four designed and eight published primer sets across parasite groups using 22 gDNA isolates from single eukaryotic endoparasites as templates for PCR. For single organisms used for DNA extraction (parasitic worms and protozoa), samples (whole adult worms, cysts, proglottids, axenic cultures, or purified DNA) were obtained from expert parasitologists or from reputable reagent repositories (for sample details including sources see Supplementary Table 3). 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  $-20^{\circ}\text{C}$ . PCR conditions were as follows: 1 X Platinum II Hot Start PCR MasterMix (ThermoFisher, Waltham, Massachusetts, USA), 0.2  $\mu\text{M}$  forward primer with Nextera adapter, 0.2  $\mu\text{M}$  reverse primer with Nextera adapter, ThermoFisher 0.2 X Platinum II GC Enhancer, 0.8 ng/ $\mu\text{l}$  gDNA in a total 12.5  $\mu\text{l}$  reaction;  $94^{\circ}\text{C}$  for 2 min, 30 cycles of [ $94^{\circ}\text{C}$  for 15 s,  $60^{\circ}\text{C}$  for 15 s,  $68^{\circ}\text{C}$  for 15 s], and hold at  $4^{\circ}\text{C}$ . 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<sup>89</sup>. We addressed this issue by extracting, amplifying, and cloning parasite DNA from 16 vouchered parasite specimens from verified sources or identified by experts. 18S rRNA sequences were amplified with full-length universal or group-specific primers (see Supplementary Tables 1, 4) 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 (pDNA) extracted from verified transformants was mixed at equimolar ratios to create the equimolar EukMix community standard reagent (Supplementary Table 4). 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<sup>90</sup>. The actual distribution of abundances of eukaryotic endosymbionts in vertebrate hosts, however, is not even<sup>91</sup>. For example, mathematical models indicate that communities of vertebrate gut parasites are best represented by a logarithmic function<sup>92</sup> with a fold difference of up to  $\sim 150$  between the most abundant and least abundant species<sup>93</sup>. Staggered communities have also been shown to test the dynamic range and detection limits of high-throughput assays better than even communities<sup>94,95</sup>. We

therefore created a staggered mock community with concentrations of 18S gene copies following a logarithmic function (Log EukMix) with the same 16 organisms as the Equimolar EukMix, with a maximum 400-fold difference among organism 18S abundances (Supplementary Table 4).

Metabarcoding using new and published primer sets was performed in triplicate with Equimolar EukMix and Log EukMix community standards 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 communities 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 Pearson Correlation Coefficients in R v.3.6.3, and GraphPad Prism v.8.4.3 was used for graphing data and for statistical analyses. The bubble plot representing percent abundances of all EukMix components was created with the bubble.pl script<sup>96</sup>.

### Testing VESPA for off-target amplification of prokaryotic, fungal, and host sequences

In addition to containing eukaryotic endosymbionts of interest, fecal samples contain bacteria, fungi, and host material<sup>38,64</sup>, and issues with abundant off-target reads originating from these organisms have been reported with 18S primers<sup>97,98</sup>. Many 18S primer sets have particularly high coverage of fungi due to the broad targeting of eukaryotes (Table 2, final column). We assessed our VESPA primers and metabarcoding protocol for amplification of off-target organisms (i.e., bacteria, archaea, fungi, and host) using DNA extracted from whole, unprocessed human fecal samples from western Uganda that were excess material collected as part of a concluded study<sup>99</sup> ( $n=40$ ). Appropriate IRB approval (The University of Wisconsin Madison IRB #2013-1463) was obtained by collaborators prior to collection and all samples were completely de-identified before use. Samples stored in 1:1 RNAlater nucleic acid preservation solution (ThermoFisher); fecal material by volume were thawed on ice and homogenized by vortexing. 0.2 g of homogenate was added to bead beating tubes (for a total of 0.1 g fecal material per extraction) and processed using the Qiagen DNeasy PowerLyzer PowerSoil kit according to manufacturer's recommendations. gDNA was eluted in Qiagen C6 buffer and stored at  $-20^{\circ}\text{C}$ . VESPA metabarcoding and data analysis were performed as described below.

### 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 in Bolivar State, on the southern Venezuelan border with Brazil ( $n=12$ ). Non-human primate fecal samples had been collected from semi-free ranging Nigerian red capped mangabeys (*Cercocebus torquatus*) in a sanctuary<sup>100</sup> ( $n=40$ ). 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. No demographic or identifying information was provided.

**Microscopy.** Microscopic analyses of non-human primate and human feces were performed as previously described<sup>101</sup>. Briefly, one gram of formalin preserved feces was concentrated via formalin-ethyl acetate sedimentation<sup>100</sup> and the sediment was examined in its entirety at  $\times 10$  objective light magnification for gastrointestinal parasites by an expert parasitologist. Additionally, one drop of sediment from each sample was examined at  $\times 40$  objective light magnification for identification of protozoa.



**Genomic DNA isolation.** Human fecal samples were processed to remove bacteria and debris as previously described<sup>102</sup>. Note that processing fecal samples in this manner is not necessary for VESPA metabarcoding and that we recommend use of whole fecal samples in our VESPA protocol (see Supplementary Note 1–VESPA protocol). Due to the unique and valuable nature of these samples (from a remote population known to harbor a markedly diverse gut eukaryotic community and previously examined by expert parasitologists) and lack of access to whole feces from this population, we used these samples for metabarcoding while acknowledging that the methodology may not be ideal for our intended applications. Briefly, feces were diluted in PBS (0.2M 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 1.5M sucrose solution. After centrifugation for 10 min at 1700 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 Supplementary Note 1–VESPA 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 locus-specific sequences (Supplementary Table 1 for locus-specific primer sequences) with Nextera adapter sequences added to the 5' end: F-TCGTCGGCAGCGTCAGATGTGATAAGAGACAG and R- GTCTCGTGGGCTCGGAGATGTGTA-TAAGAGACAG. 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 min, 30 cycles of [94 °C for 15 s, 60 °C for 15 s, 68 °C for 15 s], 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 Hot Start 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 min, 10 cycles of [95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s], 72 °C for 5 min, 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 ×300 cycle chemistry using default index and sequencing read primers and 10–20% PhiX.

**Data processing and Bioinformatics.** We processed reads from our VESPA data sets with both QIIME 2<sup>103</sup> and DADA2 v.1.16.0<sup>104</sup> 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 PR<sup>2</sup> reference sequence database version 5.1.0 (<https://doi.org/10.5281/zenodo.7805244>), which contains 18S and 16S sequences at species-level resolution. For comparison we also tested two other taxonomy databases (DOI 10.5281/zenodo.1447329): 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 PR<sup>2</sup> database returned higher numbers of fully assigned ASVs. Any ASVs not assigned taxonomy using the PR<sup>2</sup> database were queried against the full NCBI nucleotide database on September 3<sup>rd</sup>, 2022 using MegaBLAST<sup>105</sup> with default parameters.

### Statistics and reproducibility

For PCR and metabarcoding experiments using purified DNA from single organisms or mock community standards, investigators were blinded to primer and template DNA identity, gels were scored by two independent investigators, and each experiment was successfully repeated with the same outcomes. For metabarcoding of clinical fecal samples, investigators were blinded to host species and infection status. Fecal amplicon PCR was performed successfully in triplicate and amplicons were pooled prior to library preparation. Sample sizes for fecal experiments were chosen opportunistically based on number of samples available from previous studies and no statistical method was used to predetermine sample size. No samples and no data were excluded from analysis of any experiment. Randomization was not used in this study because all samples were used for the same analysis, which was a comparison between methods.

For mock community standard analysis, statistical significance of mean fold distances from the theoretical input were derived from two-tailed Wilcoxon matched-pairs signed rank tests and Pearson correlation coefficients were calculated for the relationship between theoretical and observed read abundances. For clinical fecal sample analysis, richness and prevalence metrics were calculated for each sample and the statistical significance of differences in these metrics was determined using two-tailed Wilcoxon matched-pairs signed rank tests.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The DNA sequencing data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject ID [PRJNA944233](https://doi.org/10.5281/zenodo.1447323). The in silico PCR (% primer coverage) data, mock community standard analysis data, off-target read abundance data, and percent read abundance data generated in this study are provided in the Source Data file. Publicly available databases used in this study are: PR<sup>2</sup> reference sequence database v 5.0.1 (DOI: 10.5281/zenodo.7805244), the SILVA v128 and v132 dada2 formatted 18S 'train sets' (<https://doi.org/10.5281/zenodo.1447329>), and the SILVA SSU rRNA v138.1 Ref NR database ([https://www.arb-silva.de/fileadmin/arb\\_web\\_db/release\\_138\\_1/ARB\\_files/SILVA\\_138.1\\_SSURef\\_](https://www.arb-silva.de/fileadmin/arb_web_db/release_138_1/ARB_files/SILVA_138.1_SSURef_)



NR99\_12\_06\_20\_opt.arb.gz) licensed under Creative Commons Attribution 4.0 (<https://creativecommons.org/licenses/by/4.0/>). Source data are provided with this paper.

## References

- Whipps, J. M., Lewis, K., Cooke, R. C. In *Fungi in biological control systems* (ed Burge, M. N.) Ch. 9, 161–187 (Manchester University Press, 1988).
- Konopka, A. What is microbial community ecology? *Isme J.* **3**, 1223–1230 (2009).
- Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
- Pepper, J. W. & Rosenfeld, S. The emerging medical ecology of the human gut microbiome. *Trends Ecol. Evol.* **27**, 381–384 (2012).
- Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**, 1258–1270 (2012).
- Lafrest-Lapointe, I. & Arrieta, M. C. Microbial eukaryotes: a missing link in gut microbiome studies. *mSystems* **3** <https://doi.org/10.1128/mSystems.00201-17> (2018).
- Kodio, A., Menu, E. & Ranque, S. Eukaryotic and prokaryotic microbiota interactions. *Microorganisms* **8** <https://doi.org/10.3390/microorganisms8122018> (2020).
- Avramenko, R. W. et al. Exploring the gastrointestinal “nema-biome”: deep amplicon sequencing to quantify the species composition of parasitic nematode communities. *Plos One* **10**, e0143559 (2015).
- Matijasic, M. et al. Gut microbiota beyond bacteria-mycobiome, virome, archaeome, and eukaryotic parasites in IBD. *Int. J. Mol. Sci.* **21** <https://doi.org/10.3390/ijms21082668> (2020).
- Kohler, J. R., Hube, B., Puccia, R., Casadevall, A. & Perfect, J. R. Fungi that infect humans. *Microbiol. Spectr.* **5** <https://doi.org/10.1128/microbiolspec.FUNK-0014-2016> (2017).
- Tedersoo, L. et al. Best practices in metabarcoding of fungi: from experimental design to results. *Mol. Ecol.* **31**, 2769–2795 (2022).
- Vossbrinck, C. R. & Debrunner-Vossbrinck, B. A. Molecular phylogeny of the microsporidia: ecological, ultrastructural and taxonomic considerations. *Folia Parasitol. (Praha)* **52**, 131–142 (2005).
- Dobell, C. The discovery of the intestinal protozoa of man. *Proc. R. Soc. Med.* **13**, 1–15 (1920).
- 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.* **25**, 290–309 (2019).
- Ricciardi, A. & Ndao, M. Diagnosis of parasitic infections: what’s going on? *J. Biomol. Screen.* **20**, 6–21 (2015).
- Nadler, S. A. & DE León, G. P. Integrating molecular and morphological approaches for characterizing parasite cryptic species: implications for parasitology. *Parasitology* **138**, 1688–1709 (2011).
- Jackson, T. F. *Entamoeba histolytica* and *Entamoeba dispar* are distinct species; clinical, epidemiological and serological evidence. *Int. J. Parasitol.* **28**, 181–186 (1998).
- Fotedar, R. et al. PCR detection of *Entamoeba histolytica*, *Entamoeba dispar*, and *Entamoeba moshkovskii* in stool samples from Sydney, Australia. *J. Clin. Microbiol.* **45**, 1035–1037 (2007).
- Cristescu, M. E. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends Ecol. Evol.* **29**, 566–571 (2014).
- D’Amore, R. et al. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genom.* **17**, 55 (2016).
- Nilsson, R. H. et al. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat. Rev. Microbiol.* **17**, 95–109 (2019).
- Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* **4**, e6372 (2009).
- Parfrey, L. W. et al. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Front. Microbiol.* **5**, 298 (2014).
- Mann, A. E. et al. Biodiversity of protists and nematodes in the wild nonhuman primate gut. *Isme J.* **14**, 609–622 (2020).
- Maritz, J. M. et al. An 18S rRNA Workflow for characterizing protists in sewage, with a focus on zoonotic trichomonads. *Microb. Ecol.* **74**, 923–936 (2017).
- Jarman, S. N. et al. Adelie penguin population diet monitoring by analysis of food DNA in scats. *Plos One* **8**, e82227 (2013).
- Bhadury, P. & Austen, M. C. Barcoding marine nematodes: an improved set of nematode 18S rRNA primers to overcome eukaryotic co-interference. *Hydrobiologia* **641**, 245–251 (2010).
- Avramenko, R. W. et al. High species diversity of trichostrongyle parasite communities within and between Western Canadian commercial and conservation bison herds revealed by nematome metabarcoding. *Parasites Vectors* **11** <https://doi.org/10.1186/s13071-018-2880-y> (2018).
- Avramenko, R. W. et al. The use of nemabiome metabarcoding to explore gastro-intestinal nematode species diversity and anthelmintic treatment effectiveness in beef calves. *Int. J. Parasitol.* **47**, 893–902 (2017).
- Poissant, J. et al. A repeatable and quantitative DNA metabarcoding assay to characterize mixed strongyle infections in horses. *Int. J. Parasitol.* **51**, 183–192 (2021).
- Dollive, S. et al. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol.* **13**, R60 (2012).
- Krogsgaard, L. R. et al. Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome. *Clin. Transl. Gastroenterol.* **9**, 161 (2018).
- Gogarten, J. F. et al. Metabarcoding of eukaryotic parasite communities describes diverse parasite assemblages spanning the primate phylogeny. *Mol. Ecol. Resour.* **20**, 204–215 (2020).
- Lamb, P. D. et al. How quantitative is metabarcoding: a meta-analytical approach. *Mol. Ecol.* **28**, 420–430 (2019).
- Sergaki, C. et al. Developing whole cell standards for the microbiome field. *Microbiome* **10**, 123 (2022).
- Marquina, D., Andersson, A. F. & Ronquist, F. New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Mol. Ecol. Resour.* **19**, 90–104 (2019).
- Bradley, I. M., Pinto, A. J. & Guest, J. S. Design and evaluation of illumina miSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl. Environ. Microbiol.* **82**, 5878–5891 (2016).
- Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
- Obiol, A. et al. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol. Ecol. Resour.* **20**, 718–731 (2020).
- Morales-Cruz, A. et al. Profiling grapevine trunk pathogens in planta: a case for community-targeted DNA metabarcoding. *BMC Microbiol.* **18**, 1–14 (2018).
- Vasselon, V. et al. Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods Ecol. Evol.* **9**, 1060–1069 (2018).
- Beermann, A. J., Werner, M. T., Elbrecht, V., Zizka, V. M. A. & 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.* **750**, 141969 (2021).
43. Bohmann, K. et al. Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Mol. Ecol. Resour.* **22**, 1231–1246 (2022).
44. Song, F., Kuehl, J. V., Chandran, A. & Arkin, A. P. A simple, cost-effective, and automation-friendly direct PCR approach for bacterial community analysis. *mSystems* **6**, e0022421 (2021).
45. 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.* **6**, 1809–1824 (2016).
46. Krehenwinkel, H. et al. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci. Rep.* **7**, 17668 (2017).
47. Deagle, B. E. et al. Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Mol. Ecol.* **28**, 391–406 (2019).
48. Burki, F., Sandin, M. M. & Jamy, M. Diversity and ecology of protists revealed by metabarcoding. *Curr. Biol.* **31**, R1267–R1280 (2021).
49. Macherioutou, L. et al. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecol. Evol.* **9**, 1211–1226 (2019).
50. Giebner, H. et al. Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. *Mol. Ecol. Resour.* **20**, 1333–1345 (2020).
51. Leray, M. & Knowlton, N. Censusing marine eukaryotic diversity in the twenty-first century. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **371** <https://doi.org/10.1098/rstb.2015.0331> (2016).
52. Mohrbeck, I., Raupach, M. J., Martínez Arbizu, P., Knebelberger, T. & Laakmann, S. High-throughput sequencing—the key to rapid biodiversity assessment of marine metazoa?. *PLoS One* **10**, e0140342 (2015).
53. Fernández, S. et al. Evaluating freshwater macroinvertebrates from eDNA metabarcoding: a river Nalón case study. *PLoS One* **13**, e0201741 (2018).
54. Robinson, C. V. et al. Multi-marker DNA metabarcoding detects suites of environmental gradients from an urban harbour. *Sci. Rep.* **12**, 10556 (2022).
55. Grey, E. K. et al. Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys. *Sci. Rep.* **8**, 8843 (2018).
56. Tytgat, B. et al. Monitoring of marine nematode communities through 18S rRNA metabarcoding as a sensitive alternative to morphology. *Ecol. Indic.* **107** <https://doi.org/10.1016/j.ecolind.2019.105554> (2019).
57. Casey, J. M. et al. DNA metabarcoding marker choice skews perception of marine eukaryotic biodiversity. *Environ. DNA* **3**, 1229–1246 (2021).
58. Zhang, G. K., Chain, F. J., Abbott, C. L. & Cristescu, M. E. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolut. Appl.* **11**, 1901–1914 (2018).
59. Pearman, J. K. et al. Metabarcoding as a tool to enhance marine surveillance of nonindigenous species in tropical harbors: a case study in Tahiti. *Environ. DNA* **3**, 173–189 (2021).
60. Ammon, U. V. et al. Combining morpho-taxonomy and metabarcoding enhances the detection of non-indigenous marine pests in biofouling communities. *Sci. Rep.* **8**, 16290 (2018).
61. Clarke, L. J., Beard, J. M., Swadling, K. M. & Deagle, B. E. Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecol. Evol.* **7**, 873–883 (2017).
62. Bourret, V., Gutiérrez López, R., Melo, M. & Loiseau, C. Metabarcoding options to study eukaryotic endoparasites of birds. *Ecol. Evol.* **11**, 10821–10833 (2021).
63. Hillis, D. M. & Dixon, M. T. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.* **66**, 411–453 (1991).
64. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* **14**, e1002533 (2016).
65. Zeineldin, M., Elolimy, A., Alharthi, A. & Abdelmegeid, M. The role of the bacteriome, mycobiome, archaeome and virome in animal health and disease. *Front. Vet. Sci.* **9**, 1130187 (2023).
66. Fujimura, K. E. et al. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat. Med.* **22**, 1187–1191 (2016).
67. Nagpal, R. et al. Gut microbiome and aging: physiological and mechanistic insights. *Nutr. healthy aging* **4**, 267–285 (2018).
68. Underhill, D. M. & Iliev, I. D. The mycobiota: interactions between commensal fungi and the host immune system. *Nat. Rev. Immunol.* **14**, 405–416 (2014).
69. Yu, Z. et al. Comparison of species-specific qPCR and metabarcoding methods to detect small pelagic fish distribution from open ocean environmental DNA. *PLoS One* **17**, e0273670 (2022).
70. Weerakoon, K. G. & McManus, D. P. Cell-Free DNA as a diagnostic tool for human parasitic infections. *Trends Parasitol.* **32**, 378–391 (2016).
71. Maldonado, A., Simoes, R. O., Luiz, J. S., Costa-Neto, S. F. & Vilela, R. V. 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.* **94**, e68 (2019).
72. Abraham, J. S. et al. Techniques and tools for species identification in ciliates: a review. *Int. J. Syst. Evol. Microbiol.* **69**, 877–894 (2019).
73. Hadziavdic, K. et al. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *Plos One* **9**, e87624 (2014).
74. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590–D596 (2013).
75. Benson, D. A. et al. GenBank. *Nucleic Acids Res* **46**, D41–D47 (2018).
76. Yilmaz, P. et al. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res* **42**, D643–D648 (2014).
77. Centers for Disease Control, G. H., Division of Parasitic Diseases and Malaria. *Alph. Index Parasitic Dis.*, <https://www.cdc.gov/parasites/az/index.html> (2020).
78. Lukeš, J., Stensvold, C. R., Jirků-Pomajbíková, K. & Parfrey, L. W. Are human Intestinal eukaryotes beneficial or commensals? *Plos Pathog.* **11**, e1005039 (2015).
79. Modrý, D., Pařčo, B., Petrželková, K. J. & Hasegawa, H. *Parasites of apes an atlas of coproscopic diagnostics*. Vol. 78 (Edition Chimaïra, 2018).
80. Taylor, M. A., Coop, R. L. & Wall, R. *Veterinary parasitology*. 4th edn, (John Wiley and Sons, Inc., 2016).
81. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
82. Tamura, K., Stecher, G. & Kumar, S. MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).
83. Ludwig, W. et al. ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
84. Riaz, T. et al. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res.* **39**, e145 (2011).

85. Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
86. Loakes, D. Survey and summary: the applications of universal DNA base analogues. *Nucleic Acids Res* **29**, 2437–2447 (2001).
87. Levin, J. D., Fiala, D., Samala, M. F., Kahn, J. D. & Peterson, R. J. Position-dependent effects of locked nucleic acid (LNA) on DNA sequencing and PCR primers. *Nucleic Acids Res* **34**, e142 (2006).
88. Brandt, B. W., Bonder, M. J., Huse, S. M. & Zaura, E. TaxMan: a server to trim rRNA reference databases and inspect taxonomic coverage. *Nucleic Acids Res.* **40**, W82–W87 (2012).
89. Wang, C. et al. 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.* **284** <https://doi.org/10.1098/rspb.2017.0425> (2017).
90. Piñol, J., Senar, M. A. & Symondson, W. O. The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Mol. Ecol. Notes* **28**, 407–419 (2018).
91. Poulin, R. *Evolutionary ecology of parasites*. (Princeton University Press, 2011).
92. Mouillot, D., George-Nascimento, M. & Poulin, R. How parasites divide resources: a test of the niche apportionment hypothesis. *J. Anim. Ecol.* **72**, 757–764 (2003).
93. Muñoz, G. & Cribb, T. H. Infracommunity structure of parasites of *Hemigymnus melapterus* (Pisces: Labridae) from Lizard Island, Australia: the importance of habitat and parasite body size. *J. Parasitol.* **91**, 38–44 (2005).
94. Ho, S. F. S., Wheeler, N. E., Millard, A. D. & van Schaik, W. Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. *Microbiome* **11**, 1–15 (2023).
95. Kleiner, M. et al. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **8**, 1558 (2017).
96. Zaikova, E. et al. Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ. Microbiol.* **12**, 172–191 (2010).
97. Kounosu, A., Murase, K., Yoshida, A., Maruyama, H. & Kikuchi, T. Improved 18S and 28S rDNA primer sets for NGS-based parasite detection. *Sci. Rep.* **9**, 15789 (2019).
98. Belda, E. et al. Preferential suppression of *Anopheles gambiae* host sequences allows detection of the mosquito eukaryotic microbiome. *Sci. Rep.* **7**, 3241 (2017).
99. Paige, S. B. et al. Combining footwear with public health iconography to prevent soil-transmitted helminth infections. *Am. J. Trop. Med. Hyg.* **96**, 205–213 (2017).
100. Friant, S., Ziegler, T. E. & Goldberg, T. L. Changes in physiological stress and behaviour in semi-free-ranging red-capped mangabeys (*Cercocebus torquatus*) following antiparasitic treatment. *Proc. R. Soc. B Biol. Sci.* **283**, 20161201 (2016).
101. Friant, S., Ziegler, T. E. & Goldberg, T. L. Primate reinfection with gastrointestinal parasites: behavioural and physiological predictors of parasite acquisition. *Anim. Behav.* **117**, 105–113 (2016).
102. Walderich, B., Müller, L., Bracha, R., Knobloch, J. & Burchard, G. D. A new method for isolation and differentiation of native *Entamoeba histolytica* and *E. dispar* cysts from fecal samples. *Parasitol. Res.* **83**, 719–721 (1997).
103. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
104. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
105. Morgulis, A. et al. Database indexing for production MegaBLAST searches. *Bioinformatics* **24**, 1757–1764 (2008).
106. Bates, S. T. et al. A preliminary survey of lichen associated eukaryotes using pyrosequencing. *Lichenologist* **44**, 137–146 (2012).
107. Bower, S. M. et al. Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. *J. Eukaryot. Microbiol.* **51**, 325–332 (2004).
108. Cavalier-Smith, T., Lewis, R., Chao, E. E., Oates, B. & Bass, D. *Helkesimastix marina* n. sp. (Cercozoa: Sainouroidea superfam. n.) a gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist* **160**, 452–479 (2009).
109. Bråte, J., Klaveness, D., Rygh, T., Jakobsen, K. S. & Shalchian-Tabrizi, K. Telonemia-specific environmental 18S rDNA PCR reveals unknown diversity and multiple marine-freshwater colonizations. *BMC Microbiol.* **10**, 168 (2010).
110. Geisen, S. et al. Integrating quantitative morphological and qualitative molecular methods to analyse soil nematode community responses to plant range expansion. *Methods Ecol. Evol.* **9**, 1366–1378 (2018).
111. Comeau, A. M., Li, W. K., Tremblay, J. E., Carmack, E. C. & Lovejoy, C. Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS One* **6**, e27492 (2011).
112. DeMone, C. et al. Application of next generation sequencing for detection of protozoan pathogens in shellfish. *Food Waterborne Parasitol.* **21**, e00096 (2020).
113. Hugerth, L. W. et al. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One* **9**, e95567 (2014).
114. Machida, R. J. & Knowlton, N. PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS One* **7**, e46180 (2012).
115. Sikder, M., Vestergård, M., Sapkota, R., Kyndt, T. & Nicolaisen, M. Evaluation of metabarcoding primers for analysis of soil nematode communities. *Diversity* **12**, 388 (2020).
116. Stoeck, T. et al. Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol.* **7**, 72 (2009).
117. Wood, J. R. DNA barcoding of ancient parasites. *Parasitology* **145**, 646–655 (2018).
118. Zhan, A. B. et al. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. *Methods Ecol. Evol.* **4**, 558–565 (2013).

## 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 (T.L.G.) and the UW Madison Parasitology and Vector Biology Training Grant T32AI007414 (L.A.O.).

## Author contributions

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

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-44521-3>.

**Correspondence** and requests for materials should be addressed to Leah A. Owens or Tony L. Goldberg.

**Peer review information** *Nature Communications* thanks John Gilleard and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024



## Supplementary Information- Tables

**Supplementary Table 1. 18S primers used in this study (locus specific sequences).**

<b>Primer Name</b>	<b>Reference</b>	<b>F/R</b>	<b>Sequence (5' – 3')</b>	<b>Region</b>
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	ACATTCTTGCAAATGCTTTCGCAG	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	TGGTGCCAGCASC CGCG	V4
21b8R	This study	R	TCAATTYCTTIAASTTTC	V4
EukA_F	Medlin 1988	F	AACCTGGTTGATCCTGCCAGT	5' terminus
EukB_R	Medlin 1988	R	TGATCCTTCTGCAGGTTACCTAC	3' terminus
1520_R	Lopez-Garcia 2003	R	CYGCAGGTTACCTAC	3' terminus
V3Mod_F	This study (modified from Flaherty 2018)	F	CCGGAGAGRGAGCMTKAG	5' terminus
EukBshort_R	This study (modified from Medlin 1988)	R	CCTTCCGCAGGTTACCTAC	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	TGGATAACTGTAATTRACTCT	5' terminus
LAOTW3R	This study	R	GACCTYACTAAACCATTCAATC	3' terminus

F, Forward primer; R, Reverse primer.

**Supplementary Table 2. 18S V4 unresolved amplicon sequences.**

Group	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6
Blastocystis (n = 141)	<i>Blastocystis hominis</i>	<i>Blastocystis pythoni</i>				
	<i>Blastocystis hominis</i>	<i>Blastocystis ratti</i>				
Ciliophora (n = 172)	<b><i>Bandia smalesae</i></b>	<b><i>Bandia tammar</i></b>	<b><i>Triplumaria sukuna</i></b>			
	<i>Triplumaria solea</i>	<i>Triplumaria dvoinosi</i>				
Apicomplexa (n = 1,476)	<i>Besnoitia oryctofelisi</i>	<i>Besnoitia darlingi</i>				
	<i>Cryptosporidium andersoni</i>	<i>Cryptosporidium muris</i>				
	<i>Cryptosporidium hominis</i>	<i>Cryptosporidium parvum</i>				
	<i>Cryptosporidium hominis</i>	<i>Cryptosporidium parvum</i>				
	<i>Cryptosporidium hominis</i>	<i>Cryptosporidium parvum</i>				
	<i>Eimeria tenella</i>	<i>Eimeria necatrix</i>				
	<i>Goussia janae</i>	<i>Goussia pannonica</i>				
	<b><i>Neospora caninum</i></b>	<b><i>Hammondia heydorni</i></b>	<b><i>Hammondia truffittae</i></b>			
	<i>Sarcocystis bovifelis</i>	<i>Sarcocystis sinensis</i>				
	<i>Sarcocystis levinei</i>	<i>Sarcocystis cruzi</i>				
	<i>Sarcocystis neurona</i>	<i>Sarcocystis ramphastosi</i>				
	<i>Sarcocystis tarandi</i>	<i>Sarcocystis elongata</i>				
	<i>Sarcocystis wobeseri</i>	<i>Sarcocystis columbae</i>				
	<i>Theileria buffeli</i>	<i>Theileria orientalis</i>				
	<i>Theileria buffeli</i>	<i>Theileria orientalis</i>				
	<i>Theileria orientalis</i>	<i>Theileria annulata</i>				
	<i>Theileria sergenti</i>	<i>Theileria buffeli</i>				
	<b><i>Theileria sp.</i></b>	<b><i>Babesia annae</i></b>				
	Amoebozoa (n = 317)	<i>Acanthamoeba rhyodes</i>	<i>Acanthamoeba royreba</i>			
Acanthocephala (n = 72)	<i>Polymorphus minutus</i>	<i>Polymorphus obtusus</i>				

	<i>Pomphorhynchus laevis</i>	<i>Pomphorhynchus tereticollis</i>				
	<i>Serrasentis nadakali</i>	<i>Serrasentis sagittifer</i>				
Platyhelminthes (n = 531)	<b><i>Carmyerius spatiosus</i></b>	<b><i>Explanatum explanatum</i></b>	<b><i>Fischoederius elongatus</i></b>	<b><i>Gastrothylax crumenifer</i></b>	<b><i>Olveria bosi</i></b>	<b><i>Orthocoelium sp.</i></b>
	<b><i>Clonorchis sinensis</i></b>	<b><i>Metorchis orientalis</i></b>	<b><i>Opisthorchis viverrini</i></b>			
	<b><i>Cotylophoron cotylophorum</i></b>	<b><i>Fischoederius cobboldi</i></b>	<b><i>Paramphistomum epiclitum</i></b>			
	<b><i>Drepanocephalus auritus</i></b>	<b><i>Drepanocephalus spathans</i></b>	<b><i>Mesorchis denticulatus</i></b>			
	<i>Schistosoma curassoni</i>	<i>Schistosoma mattheei</i>	<i>Schistosoma intercalatum</i>			
	<i>Diphyllobothrium nihonkaiense</i>	<i>Diphyllobothrium latum</i>				
	<i>Echinostoma paraensei</i>	<i>Echinostoma revolutum</i>				
	<b><i>Nanophyetus salmincola</i></b>	<b><i>Troglorematidae sp.</i></b>				
	<i>Preptetos caballeroi</i>	<i>Preptetos trulla</i>				
	<i>Procerovum varium</i>	<i>Procerovum cheni</i>				
	<i>Raillietina beveridgei</i>	<i>Raillietina australis</i>				
	<i>Saccocoelium brayi</i>	<i>Saccocoelium tensum</i>				
	<b><i>Stichorchis subtriquetrus</i></b>	<b><i>Paramphistomidae sp.</i></b>				
	<i>Trichobilharzia szidati</i>	<i>Trichobilharzia ocellata</i>				
	Nematoda (n = 460)	<b><i>Baylisascaris procyonis</i></b>	<b><i>Baylisascaris transfuga</i></b>	<b><i>Baylisascaris schroederi</i></b>	<b><i>Toxocara canis</i></b>	<b><i>Toxascaris leonina</i></b>
<i>Viannaia didelphis</i>		<i>Viannaia hamata</i>	<i>Viannaia minispicula</i>	<i>Viannaia viannai</i>		
<i>Cosmocercoides tonkinensis</i>		<i>Cosmocercoides dukae</i>	<i>Cosmocercoides pulcher</i>			
<i>Haemonchus similis</i>		<i>Haemonchus contortus</i>	<i>Haemonchus placei</i>			
<i>Philometra madaei</i>		<i>Philometra sawara</i>	<i>Philometra sciaenae</i>			
<i>Trichinella britovi</i>		<i>Trichinella murrelli</i>	<i>Trichinella nativa</i>			
<i>Trichinella britovi</i>		<i>Trichinella murrelli</i>	<i>Trichinella nativa</i>			
<i>Trichinella papuae</i>		<i>Trichinella pseudospiralis</i>	<i>Trichinella zimbabwensis</i>			
<b><i>Iheringascaris iniquies</i></b>		<b><i>Raphidascaroides brasiliensis</i></b>				



	<b>Anisakis sp.</b>	<b>Pseudoterranova decepiens</b>				
	<i>Brugia malayi</i>	<i>Brugia pahangi</i>				
	<i>Camallanus lacustris</i>	<i>Camallanus oxycephalus</i>				
	<i>Dichelyne mexicanus</i>	<i>Dichelyne robusta</i>				
	<i>Dracunculus insignis</i>	<i>Dracunculus lutrae</i>				
	<i>Gongylonema aegypti</i>	<i>Gongylonema neoplasticum</i>				
	<i>Onchocerca volvulus</i>	<i>Onchocerca cervicalis</i>				
	<i>Philometra cyprinirutili</i>	<i>Philometra ovata</i>				
	<b>Procamallanus pacificus</b>	<b>Spirocamallanus rarus</b>				
	<i>Strongyloides callosiureus</i>	<i>Strongyloides robustus</i>				
	<i>Strongyloides papillosus</i>	<i>Strongyloides venezuelensis</i>				
	<i>Toxocara canis</i>	<i>Toxocara cati</i>				
	<i>Travassostrongyl us callis</i>	<i>Travassostrongyl us orloffii</i>				
	<i>Trichuris arvicolae</i>	<i>Trichuris muris</i>				
	<i>Trichuris discolor</i>	<i>Trichuris ovis</i>				
	<i>Troglostrongylus brevior</i>	<i>Troglostrongylus wilsoni</i>				

Bolded rows, sequences assigned to more than one genus.

**Supplementary Table 3. Parasite specimens and sources.**

Organism	Sample type	Source	Catalog #
<i>Echinorhynchus salmonis</i>	Whole adult worms	UW Madison School of Veterinary Medicine, Dr. Tony Goldberg	NA
<i>Hymenolepis diminuta</i>	Whole adult worms	UW Madison School of Veterinary Medicine, Dr. Timothy Yoshino	NA
<i>Taenia hydatigena</i>	Cysts	Wisconsin Veterinary Diagnostic Lab	NA
<i>Bertiella studeri</i>	Proglottids	UW Madison School of Veterinary Medicine, Dr. Tony Goldberg	NA
<i>Schistosoma mansoni</i> Strain NMRI	DNA	BEI Resources	NR-28911
<i>Ascaris suum</i>	Whole adult worms	Wisconsin Veterinary Diagnostic Lab	NA
<i>Dictyocaulus viviparus</i>	Whole adult worms	Wisconsin Veterinary Diagnostic Lab	NA
<i>Dirofilaria immitis</i> Strain Missouri 2005	DNA	BEI Resources	NR-44348
<i>Trichinella spiralis</i>	DNA	USDA Animal Parasitic Diseases Laboratory	NA
<i>Encephalitozoon cuniculi</i> Strain CDC: V282	DNA	BEI Resources	NR-13510
<i>Entamoeba histolytica</i> Strain HK-9	DNA	BEI Resources	NR-175
<i>Balamuthia mandrillaris</i> CDC: V188	Axenic culture	BEI Resources	NR-46452
<i>Acanthamoeba</i> sp. Strain CDC: 12741:1	DNA	BEI Resources	NR-45611
<i>Naegleria fowleri</i> Strain CDC: V414	Axenic culture	BEI Resources	NR-46494
<i>Leishmania major</i> Strain NIH SD	DNA	BEI Resources	NR-48764
<i>Trypanosoma cruzi</i> Strain G	DNA	BEI Resources	NR-50238
<i>Giardia lamblia</i> Strain WB clone C6	DNA	BEI Resources	NR-15894
<i>Plasmodium falciparum</i> Strain D6	DNA	BEI Resources	MRA-398
<i>Babesia</i> sp. Strain MO1	DNA	BEI Resources	NR-50663
<i>Toxoplasma gondii</i>	DNA	UW Madison Department of Medical Microbiology and Immunology, Dr. Laura Knoll	NR-33509
<i>Cryptosporidium hominis</i> Strain TU502	DNA	BEI Resources	NR-2520
<i>Blastocystis hominis</i> Strain BT1	DNA	ATCC (American Type Culture Collection)	50608

NA, not applicable.

**Supplementary Table 4. EukMix components and full-length 18S cloning primers.**

Organism	FWD primer*	REV primer*	Equimolar EukMix %	Log EukMix %
1 <i>Echinorhynchus salmonis</i>	EukA_F	EukB_R	6.25	0.125
2 <i>Hymenolepis diminuta</i>	LAOTW2F	LAOTW3R	6.25	12.5
3 <i>Ascaris suum</i>	LAO18SF	LAO1498R	6.25	2.5
4 <i>Dirofilaria immitis</i>	LAO18SF	LAO1498R	6.25	0.25
5 <i>Trichinella spiralis</i>	V3mod_F	EukBshort_R	6.25	25.0
6 <i>Encephalitozoon cuniculi</i>	V3mod_F	LAOECR	6.25	16.5
7 <i>Entamoeba histolytica</i>	LAOEuk2F	EukB_R	6.25	12.5
8 <i>Balamuthia mandrillaris</i>	EukA_F	EukB_R	6.25	2.5
9 <i>Naegleria fowleri</i>	LAO380F	LAO1498R	6.25	0.0625
10 <i>Giardia intestinalis</i>	LAO380F	EukB_R	6.25	0.125
11 <i>Leishmania major</i>	LAOEukF	EukB_R	6.25	0.125
12 <i>Plasmodium falciparum</i>	EukA_F	EukB_R	6.25	2.5
13 <i>Babesia</i> sp. strain MO1	EukA_F	EukB_R	6.25	0.0625
14 <i>Toxoplasma gondii</i>	EukA_F	EukB_R	6.25	25
15 <i>Cryptosporidium hominis</i>	EukA_F	LAO1498R	6.25	0.25
16 <i>Blastocystis hominis</i>	LAOEukF	LAO1498R	6.25	0.25

\*See Supplementary Table 1 for primer sequences and references.

**Supplementary Table 5. EukMix community standard metabarcoding mean fold distance from theoretical: exact *P*-values.**

Primer 1	vs	Primer 2	Community Standard	<i>P</i> -value*
Owens 29F		Owens 2-2bF	Equimolar EukMix	0.029
Owens 29F		Stoack TAR	Equimolar EukMix	0.0017
Owens 29F		Hadz. 566	Equimolar EukMix	0.0006
Owens 29F		Bates 515	Equimolar EukMix	0.0003
Owens 29F		Owens 2-2bF	Log EukMix	0.7057
Owens 29F		Stoack TAR	Log EukMix	0.0063
Owens 29F		Hadz. 566	Log EukMix	0.0003
Owens 29F		Bates 515	Log EukMix	0.0034

\*Wilcoxon matched-pairs signed rank test, 2-tailed.

**Supplementary Table 6. VESPA vs microscopy MiSeq run metrics.**

<b>Library ID</b>	<b>SRA accession</b>	<b>Sample type</b>	<b>Raw reads</b>	<b>Reads post-quality filter</b>	<b>% lost in filter</b>
Human01	SAMN33744948	Human fecal	62,512	53239	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%

# VESPA Protocol

March 2023

## Contents

- 1- Starting material
- 2- gDNA extraction
- 3- 18S V4 Amplicon PCR
- 4- Amplicon cleanup
- 5- Indexing PCR
- 6- Library cleanup
- 7- Quantification and size determination
- 8- Pooling and sequencing

## 1- Starting material

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

Sample types tested:

Feces	Vomit	Stomach- contents
Intestine- tissue	Intestine- contents	Environmental
Entamoeba cysts	Whole helminths	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  $\mu$ l 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 $\mu$ l rxn. ( $\mu$ l)
2X Platinum II HotStart PCR Master Mix*	1X	6.0
10 $\mu$ M Forward primer	0.2 $\mu$ M	0.25
10 $\mu$ M Reverse primer	0.2 $\mu$ M	0.25
Platinum II GC Enhancer*	NA	2.5
Nuclease-free water*	NA	2.5
~10 ng/ $\mu$ l gDNA	0.8 ng/ $\mu$ l	1.0

\*Included in Master Mix Kit

Primers	Nextera adapter sequence <b>Locus-specific sequence</b>
Forward: 29_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG <b>AGCAGCCGCGGTAATTCC</b>
Reverse: 21b8_I_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG <b>TCCGTCAATTYCTTIAASTTTC</b>

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  $\mu$ l of 75% ethanol.
- 8- Wait >1 minute.
- 9- Remove and discard supernatant.
- 10- Add 175  $\mu$ l 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  $\mu$ l 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  $\mu$ l** 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  $\mu$ l of each library.

Measure the size of each library or a representative subset of libraries using an Agilent Bioanalyzer and 1  $\mu$ l of a 1 ng/ $\mu$ l 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.