Contiguous and complete assemblies of *Blastocystis* gut microbiome-associated protists reveal evolutionary diversification to host ecology

Abigail L. Lind[1,2], Nathan A. McDonald[2,3], Elias R. Gerrick[4], Ami S. Bhatt[5,6], Katherine S. Pollard[1,7,8]

1. Gladstone Institute for Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA.
2. School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia.
3. Department of Biology, Stanford University, Stanford, CA.
4. Department of Microbiology, University of Chicago, Chicago, IL.
5. Department of Genetics, Stanford University, Stanford, California, USA
6. Department of Medicine (Hematology, Blood and Marrow Transplantation), Stanford University, Stanford, California, USA
7. Department of Epidemiology & Biostatistics, University of California, San Francisco, CA.
8. Chan Zuckerberg Biohub SF, San Francisco, CA.

Running title: Genomics of host adaptation in *Blastocystis*

Corresponding author contact information

katherine.pollard@gladstone.ucsf.edu

1700 Owens Street, San Francisco, CA 94158

Competing interests statement

The authors declare no competing interests.

Abstract

Blastocystis, an obligate host-associated protist, is the most common microbial eukaryote in the human gut and is widely distributed across vertebrate hosts. The evolutionary transition of *Blastocystis* from its free-living stramenopile ancestors to a radiation of host-associated organisms is poorly understood. To explore this, we cultured and sequenced eight strains representing the significant phylogenetic diversity of the genus using long-read, short-read, and Hi-C DNA sequencing, alongside gene annotation and RNA sequencing. Comparative genomic analyses revealed significant variation in gene content and genome structure across *Blastocystis.* Notably, three strains from herbivorous tortoises, phylogenetically distant from human subtypes, have markedly larger genomes with longer introns and intergenic regions, and retain canonical stop codons absent in the human-associated strains. Despite these genetic differences, all eight isolates exhibit gene losses linked to the reduced cellular complexity of *Blastocystis,* including losses of cilia and flagella genes, microtubule motor genes, and signal transduction genes. Isolates from herbivorous tortoises contained higher numbers of plant carbohydrate-metabolizing enzymes, suggesting that like gut bacteria, these protists ferment plant material in the host gut. We find evidence that some of these carbohydrate-metabolizing enzymes were horizontally acquired from bacteria, indicating that horizontal gene transfer is an ongoing process in *Blastocystis* that has contributed to host-related adaptation. Together, these results highlight substantial genetic and metabolic diversity within the *Blastocystis* genus, indicating different lineages of *Blastocystis* have varied ecological roles in the host gut.

Introduction

The vertebrate gut is host to a diverse ecosystem of bacteria, archaea, viruses, and microbial eukaryotes. This diverse microbial community has broad impacts on the physiology and development of the host. Gut microbes can metabolize compounds that are not digestible by the host, providing access to previously inaccessible nutrients (Yadav et al. 2018). Gut microbes also influence the host's immune system and impact host health through their metabolic byproducts (Sommer and Bäckhed 2013). The microbiome's composition, which is influenced by diet, physiology, and the host immune system, varies substantially across vertebrate animals (David et al. 2014; Thorman et al. 2023; Godoy-Vitorino et al. 2012).

Compared to bacteria, substantially less is understood about microbial eukaryote impacts on host physiology and the gut ecosystem, despite the prevalence and diversity of gut microbial eukaryotes across vertebrates (Campo et al. 2020; Nash et al. 2017). The protist *Blastocystis* is the most prevalent eukaryotic microorganism in the human gut, found in 10-50% of industrialized populations and at higher prevalence in non-industrialized communities (Khaled et al. 2020; Scanlan and Marchesi 2008; Scanlan et al. 2014; Asnicar et al. 2021; Jinatham et al. 2021; Scanlan and Stensvold 2013). Beyond humans, *Blastocystis* colonizes a wide range of animals including other mammals, birds, reptiles, fish, amphibians, and insects (Gantois et al. 2020; Boreham and Stenzel 1993; Yoshikawa et al. 2016). *Blastocystis* is a genetically diverse genus, with strains adapting to varied gut environments across different host species (Stensvold et al. 2007). Its presence is associated with positive health outcomes in humans, including lower levels of gut inflammation and metabolic health (Piperni et al.

70 2024; Asnicar et al. 2021; Nieves-Ramírez et al. 2018). However, very little is
71 understood about how *Blastocystis* contributes to the gut environment and interacts
72 with other gut microbiota.

73 *Blastocystis* belongs to the Opalinata lineage of stramenopiles, a group of
74 protists that are obligate animal gut commensals (Kostka 2017). Unlike most
75 stramenopiles, *Blastocystis* is non-motile, lacks flagella or surface hairs, and grows
76 predominantly as a round cell with large central vacuole (Zierdt 1988). The *Blastocystis*
77 genus contains numerous genetically distinct groups categorized into subtypes based
78 on polymorphism in the 18S rRNA gene. Thus far, 40 distinct subtypes have been
79 proposed, and this number is continuing to expand (Santin et al. 2024; Stensvold and
80 Clark 2020). To date, strains from three axenic lab-cultured subtypes (ST1, ST4, and
81 ST7), all of which are subtypes observed in humans, have been sequenced,
82 annotated, and deposited in public databases (Denoeud et al. 2011; Gentekaki et al.
83 2017). While these subtypes span a narrow range of the *Blastocystis* phylogenetic
84 diversity, they have very different genomes and gene contents (Gentekaki et al. 2017).
85 These genomes have either gapped scaffolds (ST7) or highly fragmented non-
86 contiguous genome assemblies (ST1, ST4), limiting the ability to determine synteny
87 between subtypes, and two genomes were annotated without consideration for the
88 unique *Blastocystis* phenomenon where a portion of genes lack DNA-encoded
89 termination codons. Thus, we lack highly complete and well annotated genomes
90 necessary for establishing cases of gene loss, horizontal transfer, and sequence
91 evolution so that functional repertoires can be associated with strain diversification.

Like most host-associated eukaryotes, *Blastocystis* grows best in the presence of other microbiota. Establishing robust axenic cultures of intestinal protists including *Blastocystis* is laborious and often either fails to remove all microbial contaminants or results in the death of the protist (Zierdt and Williams 1974; Tan 2008). However, removing sources of microbial contamination is important for eukaryotic genome assembly via short-read sequencing (Gough et al. 2020). As a result, genome sequencing for *Blastocystis* and other intestinal protists has been limited to axenic cultures (Keeling et al. 2014; Gough et al. 2020). To overcome this barrier, we establish high quality and contamination-free genomes from xenic cultures of *Blastocystis* grown together with other microbes by using long-read Nanopore and Hi-C sequencing technologies coupled with short reads and metagenomic assembly techniques to establish genomes of *Blastocystis* strains that are grown together with other microbes. The longer read length of Nanopore sequencing helps span repetitive regions, resulting in greater contig length and fewer misassembles, and has been successfully applied to assembling bacterial metagenomes (Moss et al. 2020). Hi-C sequencing, which captures spatial interactions between DNA fragments, can further improve genome assembly by correcting misassemblies and scaffolding contigs (Šimková et al. 2024). By combining these technologies with metagenomic assembly techniques, we were able to generate nearly complete *Blastocystis* genomes without the need for axenic culturing. This approach allows us to investigate the full genetic diversity of *Blastocystis*, even in the presence of other microbes, and to assemble highly contiguous genomes that were previously difficult to obtain.

114 Here, we applied this hybrid sequencing strategy to eight *Blastocystis* strains,
115 five isolated from humans and three isolated from the herbivorous tortoise *Testudo*
116 *horsfeldii.* These isolates, five of which are xenic and three of which are axenic, span
117 the phylogenetic diversity of the *Blastocystis* genus, enabling us to investigate
118 *Blastocystis* evolution and determine how *Blastocystis* has adapted to thrive in animal
119 hosts with different diets, physiologies, and resident microbiota. Our highly contiguous
120 and complete genomes allowed us to investigate patterns of genome evolution within
121 and between subtypes, and to compare the genus to related protists. This work
122 demonstrates how high-quality genome assemblies can be created from xenic protist
123 culture by combining deep long-read sequencing with metagenomic assembly
124 techniques, enabling rigorous comparative genomic analyses.
125

126 **Results**

127 *Capturing Blastocystis diversity*

128 To investigate the evolution of the *Blastocystis* genus, from its initial divergence
129 through its later diversification, we sequenced eight diverse *Blastocystis* strains.
130 Strains were cultured anaerobically on egg media overlayed with a balanced salt
131 solution and 25% heat-inactivated horse serum. Five of these strains correspond to 3
132 different previously defined “subtypes” (genetically similar groups historically identified
133 in mammalian and avian species (Stensvold and Clark 2020). These five strains were
134 originally cultured from human fecal samples and were obtained from the American
135 Type Culture Collection. These human isolated strains correspond to three of the four
136 most prevalent *Blastocystis* “subtypes” (Beghini et al. 2017), genetically similar groups

137 historically identified in mammalian and avian species (Stensvold and Clark 2020): Two
138 subtype (ST) 1 (NandII, JDR), two ST3 (DL, NMH), and one ST4 (BT1). The remaining
139 three strains were isolated from companion animal tortoises: two from different
140 individuals of the domestic Russian tortoise species *Testudo horsfeldii* (Rus-B, Rus-S),
141 and one from a Hermann's tortoise *Testudo hermanni* (Hermann's). We performed
142 comparisons of the 18S rRNA gene of these our strains to other *Blastocystis* strains,
143 finding that the tortoise isolates were distantly related to the human isolates and do not
144 belong to a described mammalian or avian subtype (Figure 1A, Table S1). Analysis of
145 all eight cultures by light microscopy revealed predominantly vacuolar *Blastocystis*
146 cells, containing a large central vacuole and a thin peripheral ring of cytoplasm. While
147 *Blastocystis* have been observed to adopt a variety of morphologies, vacuolar
148 morphologies dominate cultured samples (Figure 1B).

149 To explore how the *Blastocystis* genus has emerged and diversified, we
150 compared the *Blastocystis* genomes generated in this work to those of two other
151 sequenced stramenopiles: *Cafeteria burkhardae*, a free-living heterotrophic marine
152 flagellate (Hackl et al. 2020), and the recently sequenced *Proteromonas lacertae*, a
153 reptile gut commensal closely related to *Blastocystis* (Záhonová et al. 2023). These
154 organisms are the closest relatives to *Blastocystis* that have sequenced genomes.
155 Genome comparisons allowed us to reconstruct three evolutionary transitions that
156 have occurred during stramenopile evolution: the transition from a free-living aquatic
157 lifestyle to an anaerobic lifestyle inside the animal gut (comparing *Cafeteria* to
158 *Blastocystis*/*Proteromonas*), the change in cell morphology from heterokont flagellate
159 to aflagellate, spherical cells (comparing *Cafeteria*/*Proteromonas* to *Blastocystis*), and

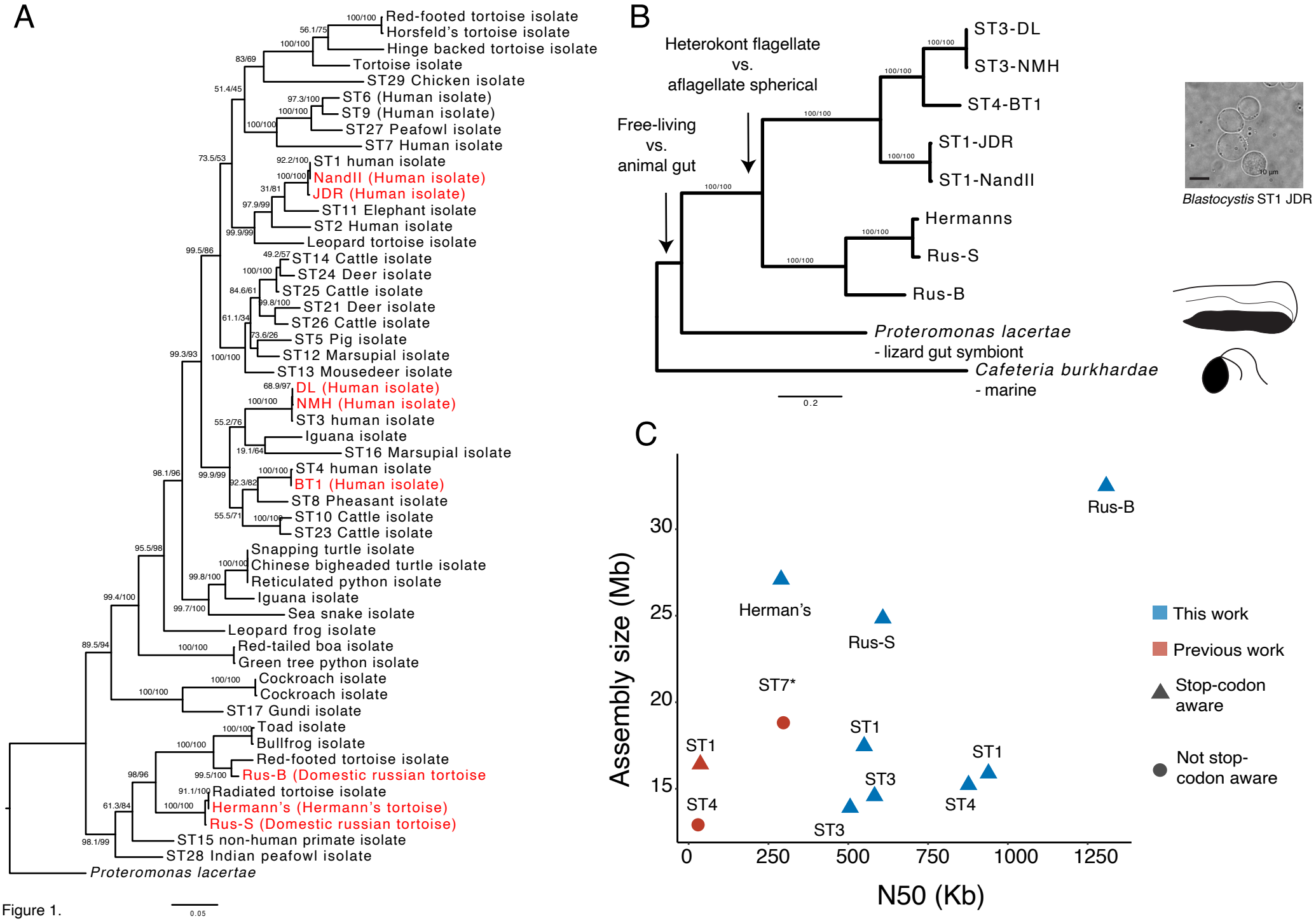

Figure 1.
(A) 18S rRNA maximum likelihood (ML) phylogeny of Blastocystis subtypes, with Proteromonas lacertae as an outgroup. Support values are 1000 replicates of ultrafast bootstrap and SH-aLRT tests (Guindon et al. 2010; Hoang et al. 2018). Strains in red were sequenced in this study. (B) 13-gene concatenation ML phylogeny for all strains and species with sequenced genomes used in this study. Important evolutionary transitions are highlighted. Images represent cellular morphology of Cafeteria burkhardae, Proteromonas lacertae, and Blastocystis.
(C) Assembly statistics of Blastocystis strains sequenced in this work compared with previously sequenced and annotated genomes. Stop-codon aware and not stop codon aware designations refer to gene annotation methodologies that take into consideration that Blastocystis genes can lack stop codons or do not consider this, respectively. The ST7 N50 is a gapped scaffold N50 generated with Sanger sequencing scaffolding.

160 adaptations to different animal hosts (comparing *Blastocystis* tortoise to *Blastocystis*

161 human) (Figure 1B).

162

163 Table 1. Genome assembly statistics for all genomes used in this study.

Species	Strain	Subtype	Source	Citation	Scaffolds (#)	Size (Mb)	N50 (Kb)	GC content (% GC)	BUSCO content[2]	Protein-coding genes
Blastocystis sp.	JDR	1	Human	**This study**	114	17.5	550	54.5%	61.2%	7891
Blastocystis sp.	NandII	1	Human	**This study**	54	16.0	939	54.7%	61.2%	7157
Blastocystis sp.	NMH	3	Human	**This study**	36	14.6	582	51.3%	58.0%	6571
Blastocystis sp.	DL	3	Human	**This study**	56	14.0	506	51.6%	55.7%	6154
Blastocystis sp.	BT1	4	Human	**This study**	34	15.3	882	39.4%	57.6%	7421
Blastocystis sp.	Rus-B	N/A	Russian tortoise	**This study**	33	33.3	1323	23.1%	43.2%	8547
Blastocystis sp.	Rus-S	N/A	Russian tortoise	**This study**	258	26.1	505	21.8%	42.0%	8501
Blastocystis sp.	Hermanns	N/A	Hermann's tortoise	**This study**	229	27.1	289	21.5%	42.4%	8349
Blastocystis sp.	WR1	4	Laboratory Wistar rat	No publication	1,301	12.9	30	39.5%	56.5%	5707
Blastocystis sp.	NandII	1	Human	Gentekaki et al. 2017	580	16.5	79	55.0%	61.1%	6544
Blastocystis sp.	Singapore	7	Human	Denoeud et al. 2011	54	18.8	901	45.0%	52.2%	6020
Proteromonas lacertae	LA	N/A	Sand lizard	Záhonová et al. 2023[2]	1,447	52.2	93	27.0%	62.7%	23189
Cafeteria burkhardae	BVI	N/A	Environmental	Hackl et al. 2020	169	36.3	465	70.0%	54.9%	8591

164

165 1. BUSCO version 5.6.1 with eukaryote_odb10 lineage, metaeuk genome mode

166 2. Gapped scaffold N50.

167 3. Citation for genome only. Annotations not publicly available and generated for comparative genomics in this work.

168

169 *Contiguous and complete Blastocystis genomes*

170 We prepared high quality, high molecular weight DNA, as well as RNA, from the

171 eight *Blastocystis* cultures. DNA was sequenced with Nanopore long-read technology.

172 Short-read DNA and RNA sequencing was also performed for assembly polishing and

173 gene annotation. Hi-C data was additionally generated to help resolve scaffolds for one
174 strain (ST4-BT1). Non-eukaryotic sequence was detected via multiple methods and
175 excluded from final assemblies (see Methods, Genome assembly and
176 decontamination). We additionally validated that our genomes were not contaminated
177 by bacterial sequence by determining that genes with high sequence similarity to
178 bacterial genes arose from polyadenylated eukaryotic transcripts (Table S2) (see
179 Methods, Gene contamination and horizontal gene transfer screens).

180 This combination of sequencing technologies enabled us to generate highly
181 contiguous genomes for all six subtypes. The N50 of these assemblies ranged from
182 289 Kb to 1.3 Mb, achieving a much higher assembly quality than currently available
183 *Blastocystis* genome datasets in NCBI (median: 37 Kb) (Table 1, Figure 1C). We used
184 the BUSCO pipeline to estimate genome completeness based on presence and
185 absence of single-copy universal eukaryotic genes (Manni et al. 2021). BUSCO
186 completeness estimates ranged from 42% to 61.2%, similar to BUSCO completeness
187 results of previously sequenced *Blastocystis* isolates and related species (Table 1,
188 Figure 1C). These values likely underestimate the true completeness, because BUSCO
189 is known to have poorer performance in protists relative to plants, animals, and fungi
190 (Saary et al. 2020). Nonetheless, the consistent recovery of a similar fraction of BUSCO
191 genes across stramenopile genomes indicates comparable completeness. It is
192 therefore striking that the sequenced *Blastocystis* strains showed substantial variability
193 in genome size, from 14.0 to 32.2 Mb. Strains in the same subtype exhibited similar
194 lengths: ST3 isolates were 14.0 and 15.3 Mb, and ST1 isolates were 17.2 and 16.0 Mb

195 in size. In contrast, the genomes of the three tortoise isolates were ~1.5 to 2 times
196 larger than human isolates, ranging from 26.0 Mb to 33.3 Mb.
197 Previous genome sequencing studies have raised questions about the ploidy of
198 *Blastocystis* and found evidence of segmentally duplicated blocks of genes. With the
199 highly contiguous genomes and multiple isolates of individual subtypes here, we
200 investigated questions of ploidy and conservation of synteny both within *Blastocystis*
201 subtypes and across the genus. *K*-mer frequency-based analysis using GenomeScope
202 (Ranallo-Benavidez et al. 2020) demonstrated a strong diploid signal in all *Blastocystis*
203 isolates, with limited evidence for aneuploidy or polyploidy. We found that strains
204 within ST3 and within ST1 were highly syntenic (Figure S1). However, synteny between
205 subtypes was much lower, broadly following the relatedness between strains based on
206 18S sequences.
207 We used a hybrid method of *de novo* gene prediction and RNA transcript
208 evidence to annotate *Blastocystis* genes (see Methods, Genome assembly and
209 decontamination). Because a substantial fraction of genes in *Blastocystis* ST1 and ST7
210 do not have DNA-encoded stop codons (Gentekaki et al. 2017; Klimeš et al. 2014)
211 (discussed below), open reading frame based gene predictions from DNA sequence
212 are inaccurate. Leveraging our RNA sequencing data, we were able to generate a
213 comprehensive gene annotation for each *Blastocystis* genome that was able to
214 accurately call genes lacking stop codons (see Methods, Gene annotation). Similar
215 approaches were taken for one previously sequenced ST1 genome (Gentekaki et al.
216 2017), but published *Blastocystis* genomes for ST4 and ST7 were not annotated with
217 these approaches, leading to incorrect gene annotations. For the *Blastocystis* strains

218 sequenced in this study, the number of genes per *Blastocystis* genome sequenced in
219 this work ranged from 6,154 to 8,547 (Table 1). Protein amino acid identity (AAI) across
220 753 single-copy orthologs was almost identical between strains of the same subtype
221 (99%-100% AAI for both ST1 and ST3), and decreased to 65-75% between human
222 *Blastocystis* isolates from different subtypes and to 40-42% between human and
223 tortoise isolates (Figure S2). Tortoise isolates had lower pairwise AAI than human
224 isolates, with the most distantly related strains in the tortoise-associated clade having
225 a protein identity of 56% AAI for the same 753 orthologs, suggesting substantial
226 genetic distance within this group.
227
228 *Genomic diversity within Blastocystis*
229 When we compared sequenced *Blastocystis* genomes, we found substantial
230 variation in the genomes between human *Blastocystis* isolates and the tortoise-
231 associated clade, including genome size. While the tortoise *Blastocystis* isolate
232 genomes are ~1.5x-2.4x larger than the human isolate genomes, this difference is not
233 reflected in protein-coding gene content, suggesting that non-genic regions are driving
234 the difference in genome size. We sought to determine why the genomes of
235 *Blastocystis* isolates in the tortoise-associated clade were larger than the genomes of
236 the *Blastocystis* human isolates.
237 We found little difference in transposable element presence across *Blastocystis*,
238 which ranges from 4.2%-6.7% of the genome in the human isolates and 5.6%-7.5% of
239 the genome in the tortoise isolates. However, we did observe an increase in low-
240 complexity regions in the tortoise isolates, comprising 0.02%-1.4% of the human

241 isolate genomes and 11.8%-12.7% of the tortoise isolate genomes (Table S3). This
242 increase in low complexity regions does not explain the difference in genome size; if
243 these regions are removed, the tortoise isolates remain 1.3x-2.1x larger than the
244 human isolates. These low complexity regions may be related to the overall lower GC
245 content of the tortoise isolates, which ranges from 21-23% in tortoise isolates to 39-
246 55% in human isolates (Table 1).

247 As repetitive elements did not explain differences in genome size, we
248 investigated differences in other genomic features. We found that the almost all introns
249 in human ST1, ST3, and ST4 *Blastocystis* are exactly 30 bp long, in agreement with
250 previous estimates of intron size in *Blastocystis* ST1 and ST7 (Gentekaki et al. 2017;
251 Denoeud et al. 2011). However, introns in the tortoise isolates are longer and more
252 variable in length (Figure 2A). We additionally found that the intergenic spaces between
253 genes are much longer in the tortoise *Blastocystis* isolate than in human isolates
254 (Figure 2B). The median intergenic length in the tortoise isolates ranges from 645-719
255 bp, while the median intergenic length in the human isolates ranges from 137-186 bp.
256 Intergenic lengths in the human isolates can frequently be extremely small, only
257 spanning tens of base pairs (Figure 2B). Together, the total amount of DNA spanned by
258 introns and intergenic regions is 13.9-16.5 Mb in the tortoise *Blastocystis* isolates,
259 whereas these regions span only 3.7-5.7 Mb in human *Blastocystis* isolates.
260 Considered together, the larger introns and intergenic distance in the tortoise isolates
261 explain the difference in their genome sizes relative to the human isolate.
262

263 *Loss of stop codons is not universal in* Blastocystis

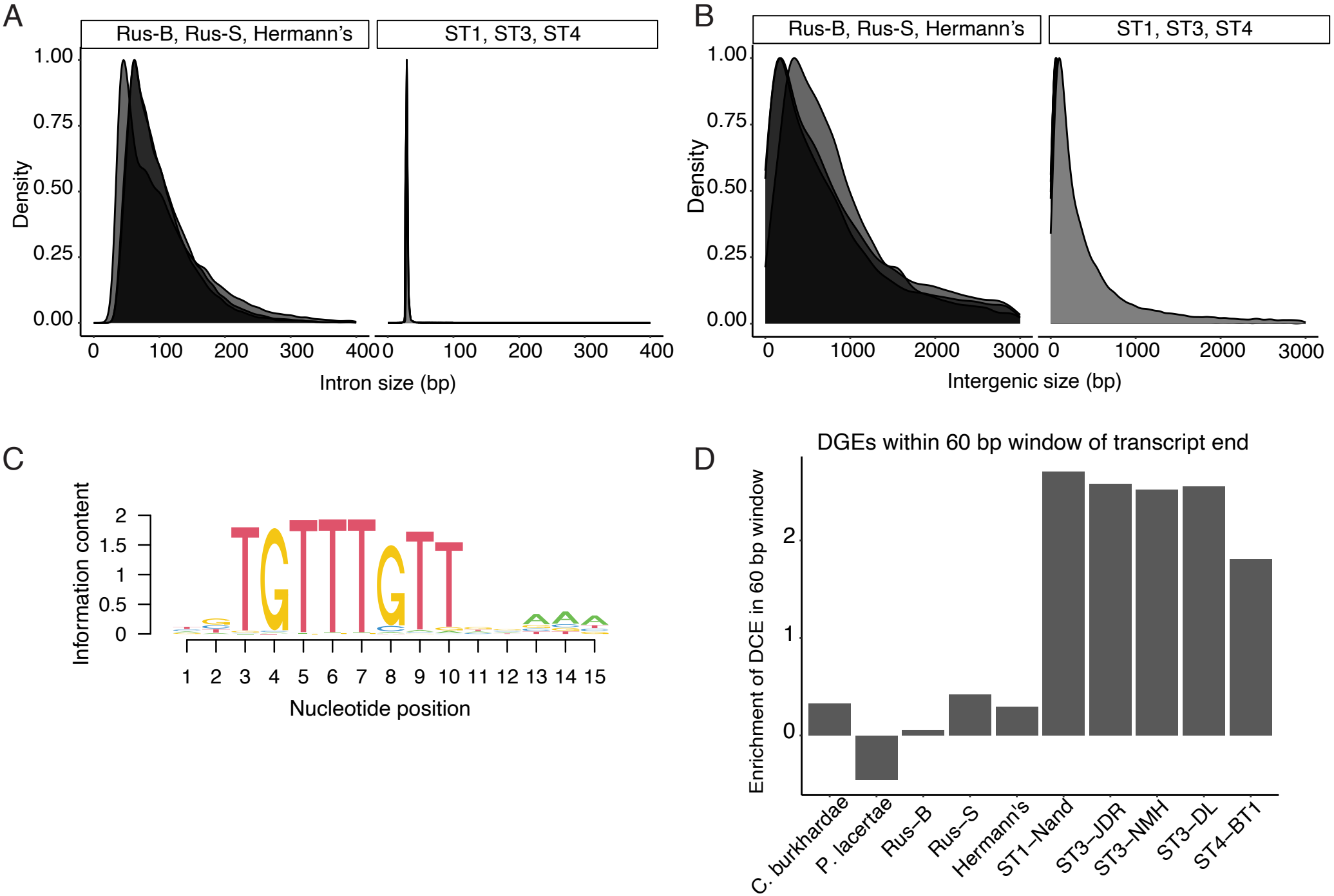

Figure 2. (A) Intron lengths and (B) Intergenic lengths in tortoise isolates (left) and human isolates right). (C) Sequence motif enriched adjacent to transcript ends in ST1, ST3, and ST4. Motif generated with STREME (Bailey 2021). This motif is referred to as "downstream gene element" (DGE). (D) Enrichment (log(observed/expected)) of transcripts with downstream conserved element within 30 bp upstream or downstream of transcript end, normalized by expected DGE occurrence given GC content.

264 In addition to genome size and gene density, we found surprising differences in
265 how genes are encoded in the genome across *Blastocystis*. In *Blastocystis* ST1 and
266 ST7, a fraction of protein coding genes (15% for ST7 and 26% for ST1) lack canonical
267 DNA-encoded termination codons (Gentekaki et al. 2017; Klimeš et al. 2014). A
268 proposed mechanism for translation of these genes is that the translational machinery
269 instead uses a U or a UG sequence in mRNA, along with one or two adenines from the
270 poly-A tail, to generate UAA and UGA on-the-fly stop codons (Klimeš et al. 2014).
271 Additionally, these genomes have a highly conserved motif (downstream gene element,
272 or DGE) roughly 5 base pairs downstream of transcription stop sites of most genes.
273 These sequences are hypothesized to direct 3' mRNA cleavage and polyadenylation
274 factors to the pre-mRNA (Gentekaki et al. 2017; Klimeš et al. 2014). In a scenario where
275 the DGE is sufficient for 3' mRNA cleavage, the loss of DNA-encoded stop codons is
276 driven by relaxation of the selection pressure maintaining them.
277 We investigated all *Blastocystis* isolates for the DGE motif, finding enrichment
278 downstream of protein coding genes in human-isolated *Blastocystis* strains, but not
279 downstream of tortoise-isolated *Blastocystis* strains or other stramenopiles (Figure 2C-
280 D). We additionally found no evidence of transcripts lacking stop codons in the
281 tortoise-associated *Blastocystis* clade, suggesting that the unique transcriptional
282 cleavage machinery described in *Blastocystis* ST1, ST4, and ST7 (Klimeš et al. 2014;
283 Gentekaki et al. 2017) evolved after the divergence of the tortoise- and the human-
284 associated clades sequenced here.
285
286 *Gene losses related to cell morphology and signaling unite the* Blastocystis *genus*

287 Despite the substantial variation in genome diversity for all *Blastocystis* strains
288 sequenced here, the strains share the typical spherical, non-motile *Blastocystis* cell
289 morphology lacking flagella. We find that all *Blastocystis* isolates are united by
290 substantial gene loss related to this morphological change across multiple functional
291 categories. All *Blastocystis* isolates have lost nearly all cilia- and flagella- related genes
292 (Figure 3, Table S4-5). They have also undergone substantial reduction of dynein and
293 kinesin microtubule motor proteins, retaining what is likely the minimal amount of
294 motor proteins required for intracellular transport. This agrees with recent analyses
295 showing reduction in these gene families in human-associated *Blastocystis* subtypes
296 (Záhonová et al. 2023), and extends that observation by demonstrating that these
297 losses unite the genus. In addition, we find that *Blastocystis* isolates have lost
298 nucleotide cyclases and many genes related to signal transduction (Figure 3, Table S4-
299 5). As cilia are major mediators of cell signaling that use cyclic nucleotides as
300 intermediaries (Johnson and Leroux 2010), the loss of cilia in *Blastocystis* likely
301 corresponds to the loss of nucleotide cyclases and reduction in signal transduction
302 related genes. Thus, our sequencing of diverse *Blastocystis* strains provides strong
303 evidence that the flagella-free morphology and associated gene losses evolved in the
304 common ancestor of the *Blastocystis* genus.
305
306 *Lineage-specific genes reveal specialization within the* Blastocystis *genus*
307 We find multiple differences in gene content in between *Blastocystis* isolates
308 that likely reflect their ecological niche. All strains in the herbivorous tortoise-
309 associated *Blastocystis* clade have a larger number of genes involved in plant

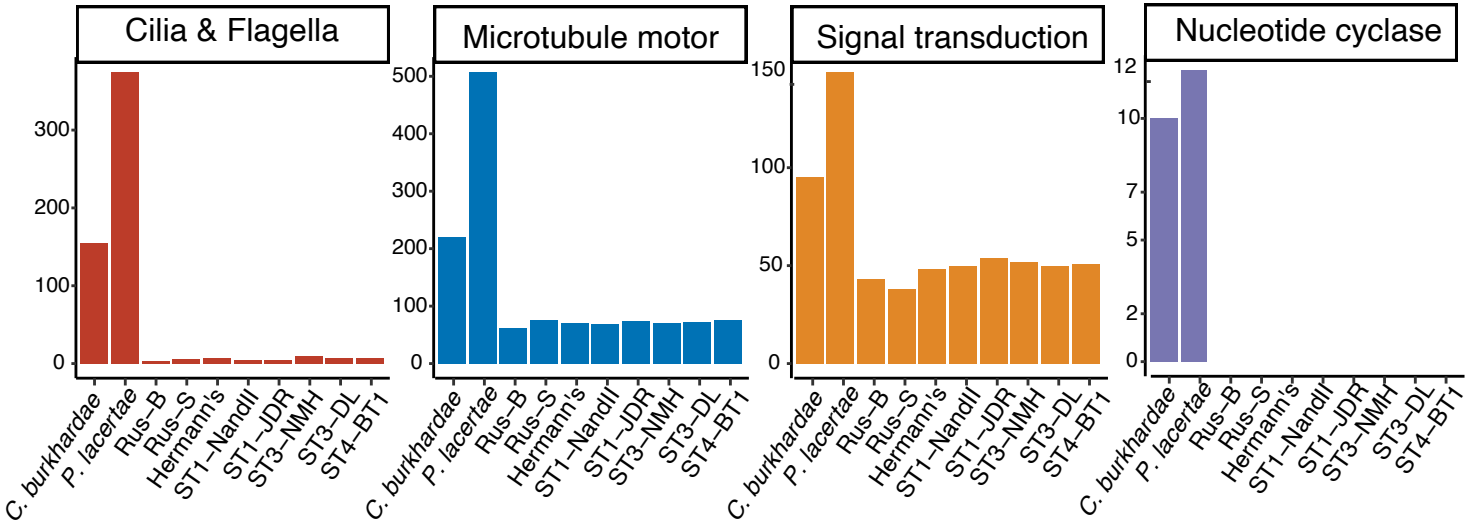

Figure 3. Gene families lost or reduced in all Blastocystis strains.
See Table S2 for GO categories and InterPro domains used for defining categories.

310 carbohydrate degradation than the human-associated subtypes. We found five
311 carbohydrate-degrading enzyme families that were either uniquely present in the
312 tortoise isolates or in higher copy numbers, four of which are specific to degrading
313 plant carbohydrates. This includes polygalacturonases, β-D-xylosidases, α-L-
314 arabinofuranosidases, α-L-fucosidases, α-L-rhamnosidases, arabinogalactanases, and
315 β-hexosaminidases (Figure 4). Polygalacturonase is a pectinase that degrades plant
316 cell walls; it was originally characterized in plants, but is also present in
317 phytopathogenic microbes and in bacterial members of ruminant microbiomes (De
318 Lorenzo and Ferrari 2002; Deng et al. 2023). β-D-xylosidases and α-L-
319 arabinofuranosidases are both critical enzymes for degrading xylan, a carbohydrate
320 specific to plant hemicellulose (Yang et al. 2004; Lagaert et al. 2014). The GH154
321 arabinogalactanase gene family that is present only in the tortoise-associated clade
322 was recently characterized in the gut bacteria *Bacteroides thetaiotamicron* as acting on
323 plant-derived arabinogalactan glycans (Cartmell et al. 2018). Finally, α-L-
324 rhamnosidases are active on a wide variety of polysaccharide residues produced by
325 plants and bacteria (Guillotin et al. 2019). Together, the presence of multiple plant-
326 specific carbohydrate digesting enzymes in the herbivorous tortoise-associated clade
327 strongly suggests that, like the bacterial microbiota, *Blastocystis* utilize dietary plant
328 fiber as a carbon source and participate in the hindgut plant fiber fermentation that is
329 characteristic of herbivorous tortoises (Stevens and Hume 1995).
330
331 *Ongoing horizontal gene transfer from bacteria in the Blastocystis genus*

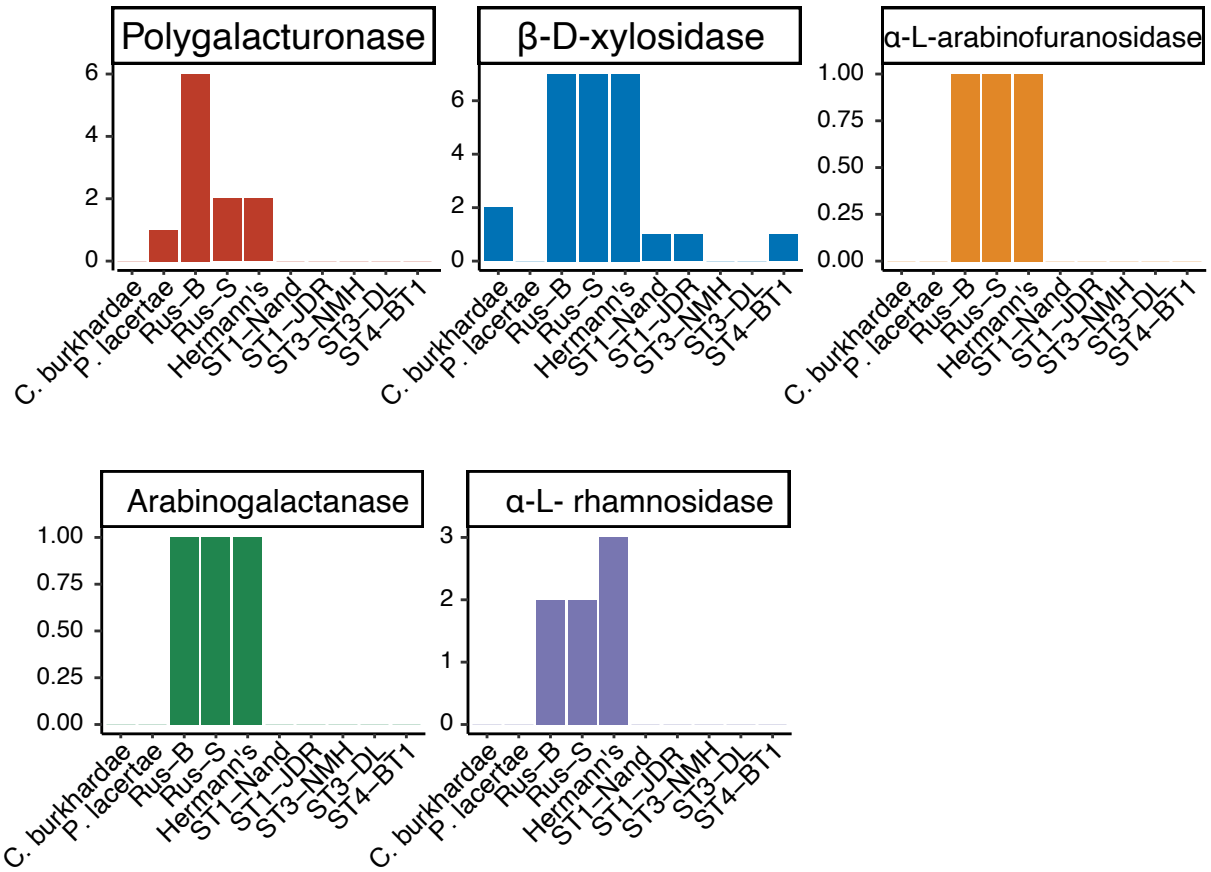

Figure 4. Clade-specific gene family expansion across *Blastocystis*, identified with InterProScan.

332 Horizontal gene transfer from bacteria has played an important role in
333 *Blastocystis* evolution, particularly with respect to survival in an anaerobic environment
334 (Eme et al. 2017). With high-quality genomes of *Blastocystis* isolates that span the
335 genus, in combination with the genomes of other stramenopiles, we can estimate times
336 of these transfer events. We re-analyzed HGT events into ST1 previously identified in
337 Eme et al 2017, and additionally screened all sequenced *Blastocystis* genomes for
338 evidence of additional transferred genes (see Methods, Screening for horizontally
339 transferred genes).
340 Previous analysis of *Blastocystis* ST1, ST4, and ST7 genomes identified 74
341 horizontal transfer events in *Blastocystis* ST1 from bacteria (Eme et al. 2017). We were
342 able to confidently identify orthologous genes for 67 of these genes (Table S6), of
343 which 27 genes had strong phylogenetic support for horizontal gene transfer from
344 bacteria to eukaryotes, 17 genes lacked strong phylogenetic support for horizontal
345 gene transfer, and the remaining 20 genes were ambiguous. In most ambiguous cases,
346 the *Blastocystis* genes were either sister to an entirely bacterial tree, there were too few
347 orthologs in the NR database to construct a gene tree, or the placing of the
348 *Blastocystis* sequences on the phylogeny lacked strong support.
349 We additionally screened all genomes sequenced for new cases of HGT. To
350 minimize the impact of potential genome contamination, we limited this screen to
351 genes that were expressed in our RNA-seq data, were shared in more than one
352 *Blastocystis* genome, and were flanked by unambiguously eukaryotic sequence to (see
353 Methods). We found 13 newly identified HGT cases, six which are unique to the
354 tortoise-associated *Blastocystis* lineage.

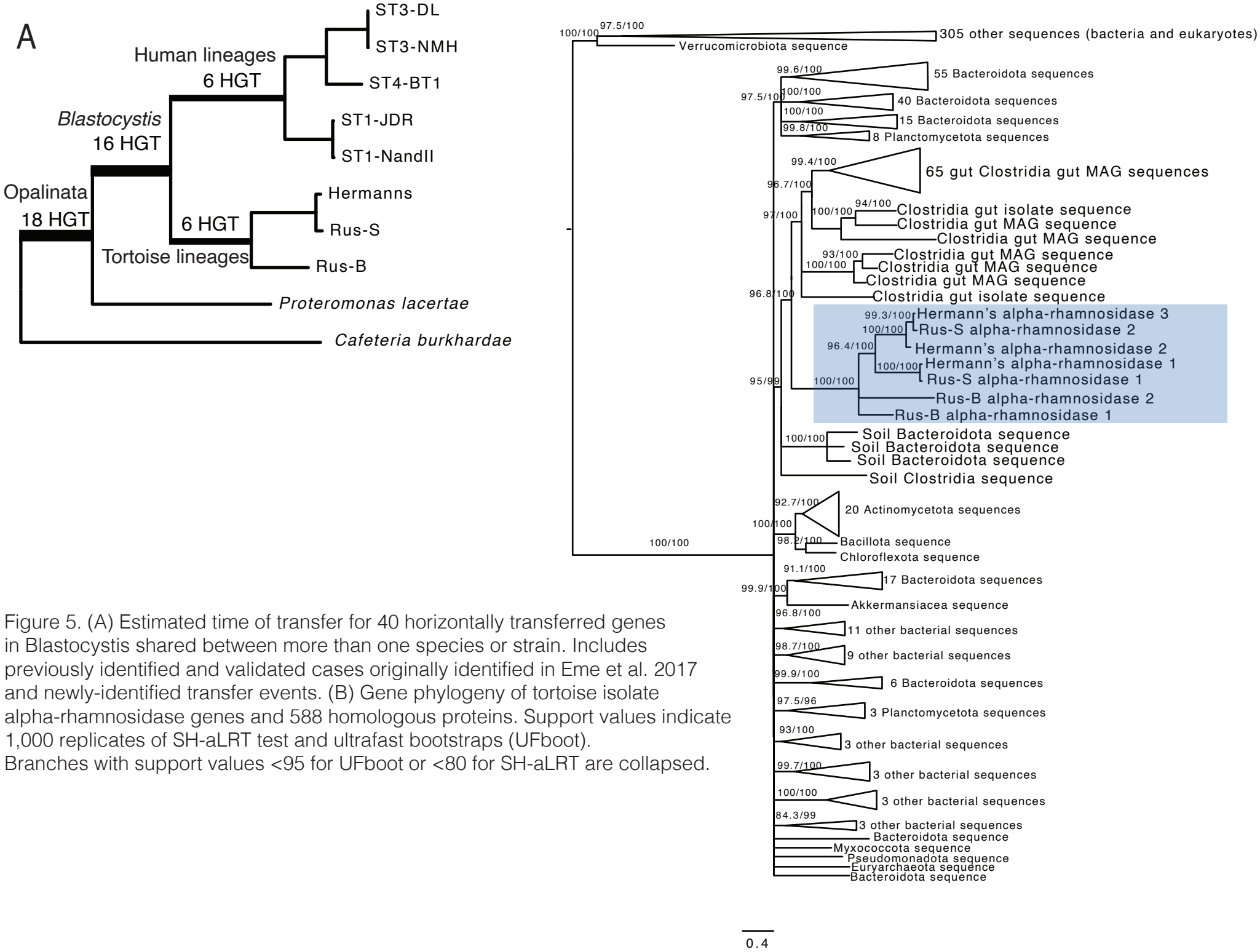

Figure 5. (A) Estimated time of transfer for 40 horizontally transferred genes in Blastocystis shared between more than one species or strain. Includes previously identified and validated cases originally identified in Eme et al. 2017 and newly-identified transfer events. (B) Gene phylogeny of tortoise isolate alpha-rhamnosidase genes and 588 homologous proteins. Support values indicate 1,000 replicates of SH-aLRT test and ultrafast bootstraps (UFboot). Branches with support values <95 for UFboot or <80 for SH-aLRT are collapsed.

355 Interestingly, two of the six tortoise-unique transfers are glycoside hydrolase
356 genes, alpha-L-rhamnosidase and arabinogalactanase, that are uniquely expanded in
357 this lineage and not present in the human *Blastocystis* isolates. The remaining transfers
358 are two GNAT family N-acetyltransferases, a cold-shock protein, and a SMI1/KNR4
359 family protein (Table S7).

360 The alpha-L-rhamnosidase gene family cleaves alpha-L-rhamnose from a large
361 diversity of molecules and is present across the tree of life. This gene is present in two
362 copies in Rus-B and Rus-S and three copies in Hermann's, and is not present in any
363 human *Blastocystis* isolates or in *P. lacertae*. Two genes with the alpha-L-
364 rhamnosidase Interpro domain were present in the *C. burkhardae* genome. We
365 constructed a gene phylogeny of the *C. burkhardae* genes and the tortoise genes
366 together with 588 homologous genes retrieved with a sequence search of the
367 *Blastocystis* and *C. burkhardae* sequences against the clustered NCBI NR database
368 (see Methods). The gene phylogeny was separated into two fully supported clades, one
369 with the *C. burkhardae* sequences, eukaryotic sequences from protists and rotifers,
370 and some bacterial sequences, and the other with bacterial sequences and the
371 *Blastocystis* sequences (Figure 5B). The tortoise sequences grouped together with
372 sequences from gut isolates and gut metagenome-assembled genomes in the
373 Clostridia bacterial class, which contains prevalent members of the animal gut
374 microbiota. Thus, the *Blastocystis* tortoise lineage likely acquired this carbohydrate
375 degrading enzyme from Clostridia bacteria present in the gut microbiome, suggesting a
376 gut-specific function.

377 The other horizontally transferred carbohydrate gene in the tortoise *Blastocystis*
378 lineage is within the recently described GH154 glycoside hydrolase gene family,
379 characterized in *Bacteroides thetaiotamicron* as acting on plant-derived
380 arabinogalactan glycans (Cartmell et al. 2018). This gene family is present exclusively in
381 the tortoise *Blastocystis* isolates and not found in any other *Blastocystis* isolates or any
382 other stramenopile genomes. A sequence search for the tortoise *Blastocystis* genes
383 against the clustered NCBI NR database did not return any eukaryotic hits. A gene
384 phylogeny of all bacterial homologs and the tortoise *Blastocystis* genes shows the
385 tortoise *Blastocystis* genes group together with bacteria in the *Bacteroidota* (syn.
386 *Bacteroidetes*) phylum (Figure S3).
387 We did not find strong evidence that the other three carbohydrate gene families
388 uniquely present or expanded in the tortoise *Blastocystis* lineage were horizontally
389 acquired from bacteria. Homologs to the polygalacturonase gene were all eukaryotic.
390 While the evolutionary history of this gene is likely complex as most homologs were
391 within Streptophyta, it was not horizontally acquired from bacteria. β-D-xylosidase and
392 α-L- arabinofuranosidase both appeared to have complex evolutionary histories but
393 without conclusive evidence of horizontal gene transfer. Specifically, several of the β-
394 D-xylosidase genes in the tortoise lineages had no eukaryotic homologs, but a gene
395 phylogeny of all β-D-xylosidase genes including the human isolates did not yield a
396 well-supported branch of the tortoise genes nested within bacteria. Similarly, the α-L-
397 arabinofuranosidase gene had both bacterial and eukaryotic homologs, but there was
398 not strong support for the placement of the *Blastocystis* sequences on this phylogeny
399 either with the eukaryotic sequences or with the bacterial sequences.

400 Together, these results indicate that horizontal gene transfer played a major role
401 in all evolutionary transitions investigated in this work: at the evolutionary transition
402 from free-living to gut symbiont, in the *Blastocystis* genus broadly, and is an ongoing
403 method of genetic diversification in this organism.
404

405 **Discussion**

406 The stramenopile protist *Blastocystis* is the most prevalent protist in the human
407 gut and is widespread throughout vertebrates. Here we present contiguous and
408 complete genomes of eight *Blastocystis* strains, five isolated from humans and three
409 from herbivorous tortoises. Long-read sequencing coupled with metagenomic
410 assembly techniques enabled us to assemble three of these genomes from xenic
411 cultures, which has been technically infeasible with standard short-read approaches.
412 Comparative genomic analyses between these *Blastocystis* genomes alongside other
413 closely related stramenopiles revealed evolutionary changes that the *Blastocystis*
414 genus has undergone to colonize a wide variety of animal guts.

415 The herbivorous tortoise isolates of *Blastocystis* diverged substantially from the
416 human isolated *Blastocystis,* and have many genomic features not seen in the human
417 isolates. We found substantial changes in genome size, with human isolates
418 possessing significantly shorter introns and intergenic regions than the tortoise
419 isolates, and substantially lower GC content in tortoise isolates. The GC content of the
420 tortoise isolates is substantially lower than the genomes of most sequenced
421 stramenopiles; of the 299 stramenopiles with sequenced genomes deposited in
422 GenBank, only three genomes have lower GC content.

423 It is not clear whether introns and intergenic regions have shrunk in size after the
424 split of the tortoise-human isolates, or if they have lengthened along the tortoise
425 lineage. Genomic streamlining is a frequently observed process in symbiotic and
426 parasitic organisms (McCutcheon and Moran 2012; Leckenby and Hall 2015) and some
427 parasitic eukaryotes have entirely or almost entirely lost introns (Wang et al. 2018;
428 Wilihoeft et al. 2001). The phenomenon of intergenic and intron size shrinking in
429 *Blastocystis* likely reflects this general process, but the unexpected differences
430 between different host-associated *Blastocystis* isolates highlights that this is an
431 ongoing evolutionary process that can proceed at different rates.
432 Despite these substantial differences in genomic streamlining, all *Blastocystis*
433 isolates sequenced here have undergone loss of the same genes associated with the
434 more complex flagellated cell morphology of other stramenopiles. *Blastocystis* isolates
435 have lost nearly all cilia- and flagella-related proteins, substantially reduced the number
436 of microtubule motor proteins, and have lost all nucleotide cyclases and many signal
437 transduction related genes. These changes unite all *Blastocystis* isolates,
438 demonstrating shared biology across the genus despite large evolutionary distances
439 and differences in genome structure between isolates.
440 We additionally find evidence consistent with host adaptation within
441 *Blastocystis*. The *Blastocystis* strain isolated from herbivorous tortoises contained
442 many carbohydrate digesting enzymes not present in other *Blastocystis* isolates (Figure
443 4). Two of these gene families were acquired via horizontal transfer from bacteria, and
444 in one case via horizontal transfer from bacteria that also colonize the animal gut. The
445 presence of plant-specific carbohydrate digesting enzymes suggests that, like some

446 ciliates and many gut bacteria, *Blastocystis* may ferment plant carbohydrates, a critical
447 element of energy homeostasis in predominantly plant-eating animals (Stevens and
448 Hume 1995).

449 While *Blastocystis* is widespread across vertebrates, there is limited evidence for
450 co-phylogeny of *Blastocystis* subtypes with their host (Figure 1). Instead, the
451 *Blastocystis* phylogeny more closely resembles that of spore-forming gut bacteria
452 (Moeller et al. 2016), with host switching as a common phenomenon. Because of the
453 lack of strong co-phylogenetic signal, open questions remain about *Blastocystis* host
454 association, including what the barriers to colonizing different hosts are and how
455 multiple strains or subtypes (commonly termed 'mixed infections') are maintained, will
456 require a denser sampling of genomes across the *Blastocystis* tree of life. As human
457 associated *Blastocystis* subtypes are not monophyletic, understanding the biology of
458 the protists that impact humans will require considering *Blastocystis* subtypes that are
459 present in the guts of diverse vertebrate hosts, and how these subtypes function in
460 different environments. Finally, the vast differences seen in the phylogenetically distinct
461 tortoise *Blastocystis* isolates sequenced here compared to previous sequenced
462 human-associated *Blastocystis* isolates highlights that substantial genetic diversity in
463 this genus has yet to be explored.

464 Five of the eight *Blastocystis* strains sequenced in this work were grown in the
465 presence of other microbes. Despite the challenges of generating highly contiguous
466 and contamination free genomes from mixed culture metagenomic sequencing, we
467 successfully combined advances in deep long read DNA sequencing and metagenomic
468 assembly techniques to acquire high-quality genomes from xenic cultures. To address

469 the challenge of annotating genes in genomes where stop codons are not universal, we
470 utilized RNA-sequencing and state-of-the-art annotation pipelines. These two technical
471 accomplishments were essential for enabling robust comparative genomic analyses of
472 Blastocystis and its closest sequenced relatives. The strategy established in this study
473 can be extended in future work to the many eukaryotic microbes that cannot be
474 cultured axenically. The resulting genomes will fill critical gaps in our understanding of
475 eukaryotic microbes and the evolution of this domain of life.
476

477 **Methods**

478 *Strain cultivation*

479 *Blastocystis* strains NandII (ATCC 50177), JDR (ATCC 50587), BT1 (ATCC
480 50608), DL (ATCC 50626), and NMH (ATCC 50754) were obtained from the American
481 Type Culture Collection. The JDR strain was labeled as subtype 3 based on rRNA
482 analysis, but 18S rRNA gene sequencing and subsequent genome sequencing
483 revealed in fact that this strain belongs to subtype 1. All strains were isolated from
484 human sources. Cultures were incubated under anaerobic conditions at 37C on
485 biphasic egg slant media overlaid with Locke's solution supplemented with 25% heat
486 inactivated horse serum (ATCC Medium 1671). Media was pre-reduced for 24 hours
487 before use. ST1 and ST4 strains were maintained in axenic culture, while ST3 and
488 tortoise isolates were grown in xenic culture alongside other microbes present in the
489 original samples. These xenic culture strains proved resistant to axenization.

490 *Blastocystis* strains from tortoises were isolated from fresh fecal material that
491 was processed within 8 hours of sampling. Fecal material was resuspended in sterile

492 PBS, strained using a 70µm cell strainer, and centrifuged at 700xg for 5 minutes at
493 room temperature. The pellet was then washed 3 times with PBS before being
494 resuspended in 80% Percoll. This was overlaid with 40% Percoll and centrifuged at
495 100xg for 10 minutes at room temperature. Protists were removed from the top of the
496 40% Percoll layer, washed in PBS, and resuspended in pre-reduced growth medium.
497
498 *DNA extraction and sequencing*
499 High molecular weight DNA was isolated from *Blastocystis* cultures with a
500 modified CTAB-based protocol (adapted from McLay 2017) that minimizes
501 polysaccharide contamination. *Blastocystis* cells were collected from cultures by
502 pelleting with centrifugation at 800x*g* for 5 minutes and snap frozen at -80˚ C. Frozen
503 pellets were gently resuspended in two volumes of a high salt CTAB lysis buffer (100
504 mM Tris pH 8, 1.2 M NaCl, 10 mM EDTA, 4% CTAB, 1% PVP-40). 200 µg/mL RNAse A
505 and 200 µg/mL proteinase K were added before incubation at 60˚ C for 90 minutes.
506 One volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added and samples
507 were mixed end over end for 15 minutes. Samples were centrifuged at 17000*g* for 10
508 minutes and the upper aqueous layer was recovered. Two additional extractions were
509 performed with pure chloroform, recovering the upper aqueous layer each time. DNA
510 was next precipitated with 2 volumes of a low salt CTAB precipitation buffer (100 mM
511 Tris pH 8, 10 mM EDTA, 2% CTAB) by end over end rotation at 60˚ C for 2.5 hours.
512 DNA was pelleted by centrifugation at 17,000x*g* for 20 minutes. DNA pellets were
513 washed twice with 75% ethanol and resuspended in 10 mM Tris pH 8. Genomic DNA
514 was further cleaned with 1.8 volumes of Ampure XP beads according to the

515 manufacturer's protocol (Beckman Coulter, Brea, CA). RNA was isolated from
516 *Blastocystis* cultures with a Trizol extraction following manufacturer's protocols
517 (Invitrogen, Waltham, MA).
518 Nanopore sequencing libraries for all human isolates and for the Rus-B strain
519 were prepared using the Genomic DNA by Ligation protocol SQK-LSK109 (Oxford
520 Nanopore Technologies, Oxford, UK) following the manufacturer's instructions,
521 sequenced on a full FLO-MIN106D R9 flow cell on a MinION sequencer, and
522 basecalled with guppy version 5.0.16. Nanopore sequencing libraries for the Rus-S and
523 Hermann's isolates were prepared using the SQK-LSK110 kit LSK109 (Oxford
524 Nanopore Technologies, Oxford, UK) and sequenced on a full R10.4.1 flow cell, and
525 basecalled with dorado version 0.5.3 (https://github.com/nanoporetech/dorado).
526 Short read Illumina DNA and RNA libraries were prepared using the KAPA
527 HyperPlus PCR-free and KAPA mRNA capture kits (Roche, Basel, Switzerland),
528 respectively, following manufacturer's instructions. RNA integrity and library fragment
529 size distribution were quantified using an Bioanalyzer 2100 (Agilent, Santa Clara, CA).
530 Illumina DNA and RNA libraries for all human *Blastocystis* isolates and for the Rus-S
531 and Hermann's tortoise isolates were sequenced on an Illumina NextSeq 500 (Illumina,
532 San Diego, CA), and libraries for the Rus-B tortoise isolate were sequenced on an
533 Illumina MiSeq v3.
534
535 *Genome assembly and decontamination*
536 Basecalled nanopore reads were assembled using Flye version 2.9.2 with the --
537 meta option enabled (Kolmogorov et al. 2019, 2020). Assembly polishing was

538 performed using four rounds of nanopore-read polishing using racon version 1.4.3
539 (Vaser et al. 2017), three rounds of Illumina DNA polishing using PolyPolish version
540 0.5.0 (Wick and Holt 2022), and 1 round of short-read polishing using POLCA as
541 implemented in MaSuRCA version 4.0.9 (Zimin et al. 2013).

542 For the ST4 strain BT1, Hi-C data was generated using Phase Genomics
543 (Seattle, WA) Proximo Hi-C 2.0 Kit, and the Phase Genomics Proximo Hi-C genome
544 scaffolding platform was used to create chromosome scale scaffolds from the hybrid
545 assembled genome. The Hi-C scaffolding resulted only in modest improvements to the
546 original assembly quality; the number of scaffolds reduced from 45 to 34, and the
547 assembly N50 improved to 882 from 876.

548 We used three different methods to separate eukaryotic, *Blastocystis*-originating
549 sequence from other microbial sequences for xenic and axenic assemblies. For xenic
550 assemblies, sequences were binned using MetaBat2 version 2.17.60 as a binning
551 algorithm, taxonomically annotated using MetaEuk taxtocontig, and additionally
552 annotated as likely eukaryotic or prokaryotic using EukRep (Kang et al. 2019; Levy
553 Karin et al. 2020; West et al. 2018). For binning, short DNA sequencing reads were
554 aligned to the polished assembly using minimap2 version 2.26-r1175 and summarized
555 using the utility jgi_summarize_bam_contig_depths (Li 2018). Metaeuk taxonomy
556 prediction was performed by running version
557 1f7d7023bdac3d46964ca380bef1b2848685501e, metaeuk easy-predict with mmseqs
558 Swiss-Prot database build and metaeuk taxtocontig with the options '--majority 0.5 --
559 tax-lineage 1 --lca-mode 2'. EukRep v0.6.6. was run in 'balanced' mode.

560 Eukaryotic contigs were identified through a union of MetaEuk contigs
561 annotated as eukaryotic, removing any not identified as eukaryotic by EukRep. We
562 found that MetaEuk was not able to identify *Blastocystis* contigs reliably at lower
563 taxonomic levels than the superkingdom Eukaryota, which we validated by running the
564 pipeline on existing NCBI genomes. Instead, any contigs that were annotated as
565 eukaryotic were considered as potentially *Blastocystis* in origin. We cross-referenced
566 these eukaryotic contigs with metabat bins, finding bins corresponding to eukaryotic
567 sequence. In two cases, eukaryotic sequences that appeared to be *Blastocystis* in
568 origin were split across two metabat bins, and these bins were combined. We removed
569 any sequences from these bins that were annotated by MetaEuk as non-eukaryotic.
570 We verified that these final cleaned bins corresponded to *Blastocystis* by BLASTing the
571 NCBI RefSeq *Blastocystis* coding gene sequences against these bins. In several cases,
572 we found bins that were annotated as eukaryotic but did not have any hits to
573 *Blastocysis* coding sequences. These bins were discarded as likely originating from
574 non-*Blastocystis* eukaryotic organisms in the culture.
575 We found that contigs containing ribosomal RNA genes were often not binned
576 together with protein coding genes, likely due to their highly repetitive nature and
577 differences in DNA coverage. We searched for contigs containing 18S rRNA genes by
578 BLASTing 18S rRNA genes against the full assembly. Contigs with full length and high
579 percent identity hits to any Blastocystis 18S genes that were also annotated as
580 eukaryotic by MetaEuk were included in final assemblies.
581 For axenic assemblies, sequenced were not binned with MetaBat2, but MetaEuk
582 and EukRep were performed as described above to identify potential contaminating

583 sequences. This resulted in two contigs being removed from the BT1 assembly, and no
584 contigs being removed from Nand or JDR assemblies.
585 Genome completeness was determined with sequenced-based and gene-based
586 approaches. For the sequence based approach, we applied a *k*-mer frequency-based
587 approach with GenomeScope v2.0 (Ranallo-Benavidez et al. 2020). Gene-based
588 completeness was determined using BUSCO v5.6.1 with the lineage dataset
589 eukaryote_odb10 in the genome mode (Manni et al. 2021).
590
591 *Annotating repetitive elements*
592 Repetitive elements in *Blastocystis* were annotated with a combination of
593 RepeatMasker and RepeatModeler using the Dfam database version 3.8 (Smit et al.
594 2013; Flynn et al. 2020; Storer et al. 2021). Repetitive elements were predicted from all
595 genomes with RepeatModeler version 2.0.5 with LTRStruct enabled. The predicted
596 transposable elements from all *Blastocystis* genomes that had a putative predicted
597 annotation were combined, and repetitive elements were tabulated in each genome
598 with a combination of the custom repeat library created with RepeatModeler and the
599 Dfam v3.8 database with RepeatMasker 4.1.7-p1.
600
601 *Gene annotation*
602 *Blastocystis* genes occasionally do not encode stop codons, which confounds
603 prediction via open reading frame and requires RNA-seq based evidence (Klimeš et al.
604 2014; Gentekaki et al. 2017). Without consideration of lack of stop codons, common
605 gene annotation errors include insertion of erroneous introns and combining adjacent

genes. To accurately annotate the *Blastocystis* genomes, we created a pipeline that used a combination of *de novo* gene prediction with Braker3 and transcriptome assembly, combining all sets of gene models with Mikado (Venturini et al. 2018; Gabriel et al. 2023).

RNA-seq reads were aligned to assemblies with STAR version 2.7.11b (Dobin et al. 2013), with a minimum of 3 supporting reads required for each junction. Transcript assemblies were generated with Stringtie version 2.2.3 (Pertea et al. 2015), which uses RNA-seq alignments against a genome to assemble transcripts. Because *Blastocystis* has very short intergenic regions, transcript assemblers can occasionally combine multiple transcripts into one. To address this, assembled transcripts were split if an exon had zero RNA-seq coverage. Assembled transcripts less than 100 bp long were discarded. To generate *de novo* gene predictions, braker version 3.0.8 was run with OrthoDB11 Eukaryota proteins as protein evidence and the aligned RNA-seq as transcript evidence (Gabriel et al. 2023; Kuznetsov et al. 2022). The Mikado transcript scoring file (available in code repository) was optimized to choose transcripts with strictly RNA-seq supported introns, but to prioritize genes containing stop codons so long as the intron boundaries were strictly supported.

Functional annotation and orthology inference

Proteins from *Cafeteria burkhardae* BVI (formerly *Cafeteria roenbergensis* BVI), and predicted proteins from *Proteromonas lacertae* were used for comparative genomics analyses (Záhonová et al. 2023; Hackl et al. 2020). The genome assembly, transcripts, and proteins for *C. burkhardae* was downloaded from GenBank under

629 accession GCA_008330645.1. The genome sequence for *Proteromonas lacertae* was
630 downloaded from GenBank under accession GCA_002245135.1. As no predicted
631 proteins or RNA-seq data was available for *P. lacertae*, genes were predicted from this
632 genome using the Braker3 pipeline, using OrthoDB11 Eukaryota protein sequences as
633 evidence. Proteins from all *Blastocystis* assemblies, *C. burkhardae*, and *P. lacertae*
634 were annotated using a combination of InterProScan version 5.66 (Jones et al. 2014) ,
635 DeepLoc2 to predict protein localization (Thumuluri et al. 2022), and dbCAN version
636 4.0.0 to annotate carbohydrate active enzymes (Zhang et al. 2018). Orthologous
637 sequences were determined using OrthoFinder2 with diamond as the search algorithm
638 and mcl as the clustering algorithm (Emms and Kelly 2019; Buchfink et al. 2015; Li et
639 al. 2003).
640 Genes related to cilia and flagella, microtubule motor proteins, signal
641 transduction, ion channels, and nucleotide cyclase were identified using either GO
642 categories or Interpro domains corresponding to these categories, and are listed in
643 Table S4. Carbohydrate-degrading enzymes were identified based on the gene
644 annotation methods mentioned above; specifically, polygalacturonase enzymes were
645 identified with the PANTHER domain PTHR33928, beta-D-xylosidase with the Interpro
646 domain IPR044993, a-L-arabinofuranosidase with the Interpro domain IPR010720, a-L-
647 rhamnosidase with the PANTHER domain PTHR34987, and the GH154
648 arabinogalactanase with the PANTHER domain PTHR35339.
649
650 *Identifying motifs adjacent to transcript ends*

651 To identify motifs enriched downstream of transcript ends, we extracted a 60 bp
652 window up and downstream of all annotated transcript ends in all *Blastocystis*
653 genomes sequenced here, as well as *P. lacertae* and *C. burkhardae*. Motifs were
654 identified separately in each genome using STREME version 5.5.7 (Bailey 2021). A
655 motif similar to the one identified by Klimeš et al 2014 was identified in both ST1, both
656 ST3, and the ST4 human isolate, but no significant and informative motifs were
657 identified in the tortoise-isolated *Blastocystis* isolates or in the *P. lacertae* or *C.*
658 *burkhardae* genomes. We then combined the sequences from all human isolates and
659 ran STREME on this combined dataset to generate a sequence motif (Figure 2).
660 To determine enrichment of this motif in all genomes, we searched all transcript
661 ends for matches to a degenerate version of the core region of the sequence motif,
662 TGTTGTT, allowing up to two mismatches. As this motif is inherently an AT-rich
663 sequence, the probability of observing it by chance is a function of the GC content of
664 the sequence that is searched. We calculated the expected probability of seeing the 7-
665 base pair in each genome given the transcript window GC content and the number of
666 transcripts, and calculated the enrichment of this motif as the log of the observed motif
667 counts divided by the background expectation value (Figure 2).
668
669 *Strain and species phylogenies*
670 The *Blastocystis* genus phylogeny was constructed using 18S rRNA sequences
671 deposited in GenBank (Table S1) along with 18S rRNA sequences extracted from each
672 *Blastocystis* assembly and from the *Proteromonas lacertae* genome as an outgroup.
673 Sequences were aligned with MAFFT v7.505 (Katoh and Standley 2013) and a

674 maximum likelihood phylogeny constructed using IQ-TREE version 2.3.5 (Minh et al.
675 2020) using with the GTR sequence model, 1,000 ultrafast bootstrap and SH-aLRT
676 replicates, and rooted on *P. lacertae*.
677 To construct a species phylogeny of all species and strains with genomic data
678 used in this study, 13 conserved protein-coding genes identified with BUSCO version
679 5.6.1 as described above were extracted from *Cafeteria burkhardae, Proteromonas*
680 *lacertae*, and all *Blastocystis* strains sequenced here. Proteins were aligned individually
681 with MAFFT v7.505 and concatenated. Best-fit models for each partition were selected
682 using a relaxed hierarchical clustering algorithm as implemented by ModelFinder in IQ-
683 TREE version 2.3.5 (Chernomor et al. 2016; Minh et al. 2020; Kalyaanamoorthy et al.
684 2017). 1,000 ultrafast bootstrap and SH-aLRT replicates were performed.
685
686 *Gene contamination and horizontal gene transfer screens*
687 *Blastocystis* genomes were screened for bacterial contamination and for
688 possible horizontal gene transfer events using a modified version of an alien index
689 score (Gladyshev et al. 2008; Wisecaver et al. 2016). The alien index score compares
690 the best bit score of a query sequence to a group of related taxa versus to all other
691 unrelated groups, and normalizes this score by the maximum possible bit score of the
692 query sequence. It is determined by the formula $AI = \frac{bbsO}{bbsS} - \frac{bbsI}{bbS}$, where bbsO is the
693 bit score of the best hit to a species outside of the group lineage, bbsI is the bit score
694 of the best hit to a species within the group lineage (skipping self hits), and bbsS is the
695 bit score of the query aligned to itself. The alien index was calculated for each

696 *Blastocystis* protein by querying proteins against the NCBI nr database using
697 DIAMOND version 2.1.6 (Buchfink et al. 2015)(code availability:
698 https://github.com/allind/alienindex). The self-group was designated as *Blastocystis*
699 (taxid: 12967), and the group lineage was set as all eukaryotes (taxid: 2759).
700 As an additional control for whether our genomes had high levels of genome
701 contamination, we determined what fraction of genes with high alien index scores were
702 likely transcribed by *Blastocystis.* Our RNA-seq strategy involved poly-A purification,
703 which removes non-polyadenylated transcripts. As polyadenylation is unique to
704 eukaryotes, this removes bacterial transcripts. We quantified gene expression using
705 kallisto version v0.48.0 (Bray et al. 2016), and determined whether genes with alien
706 index scores greater than 0.2 were expressed and therefore likely truly integrated into
707 the *Blastocystis* genomes. The number of genes with alien index scores above 0.2
708 ranged from 59-81 genes across all sequenced genomes, and greater than 85% were
709 expressed in all genomes. Genes with alien index scores greater than 0.2 that were
710 expressed at TPM > 1 and had orthologs in one or more other *Blastocystis* genomes
711 were further examined for evidence of horizontal gene transfer with gene tree
712 reconstruction.
713
714 *Gene tree reconstruction for horizontal transfer*
715 Phylogenetic trees for putative horizontal gene transfer candidates were
716 constructed from all diamond blastp hits using the NCBI NR database clustered at
717 90% sequence identity and 90% sequence length (Buchfink et al. 2015). Sequence hits
718 that were outliers in length were excluded from downstream analyses. Protein

sequences were aligned with MAFFT v7.505 (Katoh and Standley 2013), and maximum likelihood phylogenies were constructed using IQ-TREE version 2.3.5 using ModelFinder to determine the best-fit substitution model, and with 1,000 ultrafast bootstrap replicates and 1,000 replicates of the SH-like approximate likelihood ratio test (Guindon et al. 2010). Branches with SH-aLRT support less than 70 and UF-boot less than 80 were collapsed and trees were midpoint rooted before being examined for horizontal transfer.

Acknowledgements

We thank members of the Pollard lab and the Bhatt lab for helpful discussions. This work was supported NHLBI grant#R01-HL160862, Chan Zuckberg Biohub SF and Gladstone Institutes to K.S.P, by the Benioff Center for Microbiome Medicine Trainee Pilot Award and Georgia Institute of Technology to A.L.L. A.SB. is supported by NIH R01AI143757 and R01AI148623, as well as a Stand Up 2 Cancer grant and a Paul Allen Distinguished Investigator Award.

Author Contributions

Conceptualization – A.L.L, A.S.B, K.S.P.; Methodology – A.L.L., N.A.M, E.S.G, A.S.B, K.S.P.; Investigation – A.L.L, K.S.P.; Resources – A.S.B., K.S.P.; Writing – original draft – A.L.L., N.A.M, E.S.G, A.S.B, K.S.P; Writing – review & editing - A.L.L., N.A.M, E.S.G, A.S.B, K.S.P; Visualization – A.L.L.; Supervision – K.S.P., A.S.B.; Funding acquisition – A.L.L., K.S.P., A.S.B.

742 **Competing Interests**

743 The authors declare no competing interests.

744

745 **DATA ACCESS**

746 Genome assemblies and annotations are available on Genbank (Table 2) and are
747 additionally available on Figshare at
748 https://figshare.com/projects/Blastocystis_genomes/186136. As no gene annotations
749 were available for *Proteromonas lacertae*, BRAKER3 predictions were generated and
750 have been made publicly available in this repository as well. RNA sequencing for all
751 strains is available from the SRA under BioProject PRJNA1041441. Code to calculate
752 the alien index is available at https://github.com/allind/alienindex. All other code used
753 in this study is available at https://github.com/allind/blastannotate. All alignments and
754 phylogenies generated in this work are available at
755 https://doi.org/10.6084/m9.figshare.28143281.v1.

756

757 Table 2. NCBI accessions for genomes generated here.

Strain	NCBI Bioproject
Rus-B (also Blastoise)	PRJNA1084774
Rus-S	Processing on Genbank, currently available on Figshare
Hermann's	Processing on Genbank, currently available on Figshare
ST1-NandII	PRJNA1084796

ST1-JDR	PRJNA1084798
ST3-DL	PRJNA1084795
ST3-NMH	PRJNA1084799
ST4-BT1	PRJNA1084775

758

759

760 **Figure legends**

761 Figure 1. (A) 18S rRNA maximum likelihood (ML) phylogeny of Blastocystis subtypes,
762 with *Proteromonas lacertae* as an outgroup. Support values are 1000 replicates of
763 ultrafast bootstrap and SH-aLRT tests (Guindon et al. 2010; Hoang et al. 2018). Strains
764 in red were sequenced in this study. (B) 13-gene concatenation ML phylogeny for all
765 strains and species with sequenced genomes used in this study. Important
766 evolutionary transitions are highlighted. Images represent cellular morphology of
767 *Cafeteria burkhardae*, *Proteromonas lacertae*, and *Blastocystis*. (C) Assembly statistics
768 of Blastocystis strains sequenced in this work compared with previously sequenced
769 and annotated genomes. Stop-codon aware and not stop codon aware designations
770 refer to gene annotation methodologies that take into consideration that Blastocystis
771 genes can lack stop codons or do not consider this, respectively. The ST7 N50 is a
772 gapped scaffold N50 generated with Sanger sequencing scaffolding.

773

774 Figure 2. (A) Intron lengths and (B) Intergenic lengths in tortoise isolates (left) and
775 human isolates right). (C) Sequence motif enriched adjacent to transcript ends in ST1,
776 ST3, and ST4. Motif generated with STREME (Bailey 2021). This motif is referred to as

777 “downstream gene element” (DGE). (D) Enrichment (log(observed/expected)) of
778 transcripts with downstream conserved element within 30 bp upstream or downstream
779 of transcript end, normalized by expected DGE occurrence given GC content.
780
781 Figure 3. Gene families lost or reduced in all Blastocystis strains. See Table S2 for GO
782 categories and InterPro domains used for defining categories.
783
784 Figure 4. Clade-specific gene family expansions across *Blastocystis,* identified with
785 InterProScan.
786
787 Figure 5. (A) Estimated time of transfer for 40 horizontally transferred genes
788 in *Blastocystis* shared between more than one species or strain. Includes
789 previously identified and validated cases originally identified in (Eme et al. 2017)
790 and newly-identified transfer events. (B) Gene phylogeny of tortoise isolate
791 alpha-rhamnosidase genes and 588 homologous proteins. Support values indicate
792 1,000 replicates of SH-aLRT test and ultrafast bootstraps (UFboot). Branches with
793 support values <95 for UFboot or <80 for SH-aLRT are collapsed.
794
795 Figure S1. Synteny of *Blastocystis* genomes based on reciprocal best blast hits. Plot
796 created with ODP (Schultz et al. 2023).
797
798 Figure S2. Median percent identity across 753 single-copy orthologs in *Blastocystis*.
799

800 Figure S3. Gene phylogeny of tortoise Blastocystis GH154 arabinogalactanase
801 enzyme. Support values indicate 1,000 replicates of SH-aLRT test and ultrafast
802 bootstrap. Branches with aLRT support <80 and UFboot support <90 are collapsed.
803 Tree is midpoint rooted.
804
805 Table S1. GenBank accessions of 18S rRNA sequences used to construct *Blastocystis*
806 phylogeny.
807
808 Table S2. Amount and percentages of genome masked per strain by RepeatMasker.
809
810 Table S3. GO and IPR categories used for gene loss described in Figure 3.
811
812 Table S4. Genes containing InterPro domains per genome assembly.
813
814 Table S5. Percentage of high alien index genes per genome with poly-A purified RNA-
815 seq expression support.
816
817 Table S6. Validation of previously reported ST1-NandII horizontal gene transfers.
818
819 Table S7. Newly identified HGT events in *Blastocystis*.
820
821
822

823 **References**

824 Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, Leeming E, Gibson R, Le
825 Roy C, Khatib HA, et al. 2021. Microbiome connections with host metabolism and
826 habitual diet from 1,098 deeply phenotyped individuals. *Nat Med* **27**: 321–332.

827 Bailey TL. 2021. STREME: accurate and versatile sequence motif discovery. *Bioinforma Oxf*
828 *Engl* **37**: 2834–2840.

829 Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM, Segata N. 2017. Large-scale
830 comparative metagenomics of Blastocystis, a common member of the human gut
831 microbiome. *ISME J* **11**: 2848–2863.

832 Boreham PFL, Stenzel DJ. 1993. Blastocystis in Humans and Animals: Morphology, Biology,
833 and Epizootiology. In *Advances in Parasitology* (eds. J.R. Baker and R. Muller), Vol. 32
834 of, pp. 1–70, Academic Press
835 https://www.sciencedirect.com/science/article/pii/S0065308X08602067 (Accessed
836 October 24, 2023).

837 Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq
838 quantification. *Nat Biotechnol* **34**: 525–527.

839 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
840 *Nat Methods* **12**: 59–60.

841 Campo J del, Bass D, Keeling PJ. 2020. The eukaryome: Diversity and role of microeukaryotic
842 organisms associated with animal hosts. *Funct Ecol* **34**: 2045–2054.

843 Cartmell A, Muñoz-Muñoz J, Briggs JA, Ndeh DA, Lowe EC, Baslé A, Terrapon N, Stott K,
844 Heunis T, Gray J, et al. 2018. A surface endogalactanase in Bacteroides thetaiotaomicron
845 confers keystone status for arabinogalactan degradation. *Nat Microbiol* **3**: 1314–1326.

846 Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for
847 Phylogenomic Inference from Supermatrices. *Syst Biol* **65**: 997–1008.

848 David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin
849 AS, Varma Y, Fischbach MA, et al. 2014. Diet rapidly and reproducibly alters the human
850 gut microbiome. *Nature* **505**: 559–563.

851 De Lorenzo G, Ferrari S. 2002. Polygalacturonase-inhibiting proteins in defense against
852 phytopathogenic fungi. *Curr Opin Plant Biol* **5**: 295–299.

853 Deng Q, Sun X, Gao D, Wang Y, Liu Y, Li N, Wang Z, Liu M, Wang J, Wang Q. 2023.
854 Characterization of Two Novel Rumen-Derived Exo-Polygalacturonases: Catalysis and
855 Molecular Simulations. *Microorganisms* **11**: 760.

856 Denoeud F, Roussel M, Noel B, Wawrzyniak I, Da Silva C, Diogon M, Viscogliosi E, Brochier-
857 Armanet C, Couloux A, Poulain J, et al. 2011. Genome sequence of the stramenopile
858 Blastocystis, a human anaerobic parasite. *Genome Biol* **12**: R29.

859 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
860 TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl* **29**: 15–21.

861 Eme L, Gentekaki E, Curtis B, Archibald JM, Roger AJ. 2017. Lateral Gene Transfer in the
862 Adaptation of the Anaerobic Parasite Blastocystis to the Gut. *Curr Biol CB* **27**: 807–820.

863 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
864 genomics. *Genome Biol* **20**: 238.

865 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
866 RepeatModeler2 for automated genomic discovery of transposable element families. *Proc*
867 *Natl Acad Sci* **117**: 9451–9457.

868 Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. 2023.
869 BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence
870 with GeneMark-ETP, AUGUSTUS and TSEBRA. *BioRxiv Prepr Serv Biol*
871 2023.06.10.544449.

872 Gantois N, Lamot A, Seesao Y, Creusy C, Li L-L, Monchy S, Benamrouz-Vanneste S,
873 Karpouzopoulos J, Bourgain J-L, Rault C, et al. 2020. First Report on the Prevalence and
874 Subtype Distribution of Blastocystis sp. in Edible Marine Fish and Marine Mammals: A
875 Large Scale-Study Conducted in Atlantic Northeast and on the Coasts of Northern
876 France. *Microorganisms* **8**: 460.

877 Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Eliáš M, Salas-Leiva DE, Herman EK, Eme L,
878 Arias MC, Henrissat B, et al. 2017. Extreme genome diversity in the hyper-prevalent
879 parasitic eukaryote Blastocystis. *PLOS Biol* **15**: e2003769.

880 Godoy-Vitorino F, Goldfarb KC, Karaoz U, Leal S, Garcia-Amado MA, Hugenholtz P, Tringe
881 SG, Brodie EL, Dominguez-Bello MG. 2012. Comparative analyses of foregut and
882 hindgut bacterial communities in hoatzins and cows. *ISME J* **6**: 531–541.

883 Gough R, Barratt J, Stark D, Ellis J. 2020. Diversity profiling of xenic cultures of Dientamoeba
884 fragilis following systematic antibiotic treatment and prospects for genome sequencing.
885 *Parasitology* **147**: 29–38.

886 Guillotin L, Kim H, Traore Y, Moreau P, Lafite P, Coquoin V, Nuccio S, de Vaumas R,
887 Daniellou R. 2019. Biochemical Characterization of the α-l-Rhamnosidase DtRha from
888 Dictyoglomus thermophilum: Application to the Selective Derhamnosylation of Natural
889 Flavonoids. *ACS Omega* **4**: 1916–1922.

890 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
891 Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the
892 Performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

893 Hackl T, Martin R, Barenhoff K, Duponchel S, Heider D, Fischer MG. 2020. Four high-quality
894 draft genome assemblies of the marine heterotrophic nanoflagellate Cafeteria
895 roenbergensis. *Sci Data* **7**: 29.

896 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the
897 Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**: 518–522.

898 Jinatham V, Maxamhud S, Popluechai S, Tsaousis AD, Gentekaki E. 2021. Blastocystis One
899 Health Approach in a Rural Community of Northern Thailand: Prevalence, Subtypes and
900 Novel Transmission Routes. *Front Microbiol* **12**.
901 https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.746340/f
902 ull (Accessed December 13, 2024).

903 Johnson J-LF, Leroux MR. 2010. cAMP and cGMP signaling: sensory systems with prokaryotic
904 roots adopted by eukaryotic cilia. *Trends Cell Biol* **20**: 435–444.

905 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell
906 A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification.
907 *Bioinformatics* **30**: 1236–1240.

908 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
909 fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**: 587–589.

910 Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive
911 binning algorithm for robust and efficient genome reconstruction from metagenome
912 assemblies. *PeerJ* **7**: e7359.

913 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
914 Improvements in Performance and Usability. *Mol Biol Evol* **30**: 772–780.

915 Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV,
916 Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote
917 Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of
918 Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biol* **12**:
919 e1001889.

920 Khaled S, Gantois N, Ly AT, Senghor S, Even G, Dautel E, Dejager R, Sawant M, Baydoun M,
921 Benamrouz-Vanneste S, et al. 2020. Prevalence and Subtype Distribution of Blastocystis
922 sp. in Senegalese School Children. *Microorganisms* **8**: 1408.

923 Klimeš V, Gentekaki E, Roger AJ, Eliáš M. 2014. A Large Number of Nuclear Genes in the
924 Human Parasite Blastocystis Require mRNA Polyadenylation to Create Functional
925 Termination Codons. *Genome Biol Evol* **6**: 1956–1961.

926 Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J,
927 Polevikov E, Smith TPL, et al. 2020. metaFlye: scalable long-read metagenome assembly
928 using repeat graphs. *Nat Methods* **17**: 1103–1110.

929 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using
930 repeat graphs. *Nat Biotechnol* **37**: 540–546.

931 Kostka M. 2017. Opalinata. In *Handbook of the Protists* (eds. J.M. Archibald, A.G.B. Simpson,
932 and C.H. Slamovits), pp. 543–565, Springer International Publishing, Cham
933 https://doi.org/10.1007/978-3-319-28149-0_4 (Accessed July 28, 2023).

934 Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM.
935 2022. OrthoDB v11: annotation of orthologs in the widest sampling of organismal
936 diversity. *Nucleic Acids Res* **51**: D445–D451.

937 Lagaert S, Pollet A, Courtin CM, Volckaert G. 2014. β-Xylosidases and α-l-
938 arabinofuranosidases: Accessory enzymes for arabinoxylan degradation. *Biotechnol Adv*
939 **32**: 316–332.

940 Leckenby A, Hall N. 2015. Genomic changes during evolution of animal parasitism in
941 eukaryotes. *Curr Opin Genet Dev* **35**: 86–92.

942 Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene discovery,
943 and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**: 48.

944 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–
945 3100.

946 Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic
947 genomes. *Genome Res* **13**: 2178–89.

948 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and
949 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for
950 Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**: 4647–4654.

951 McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev*
952 *Microbiol* **10**: 13–26.

953 McLay T. 2017. High quality DNA extraction protocol from recalcitrant plant tissues v1.
954 https://www.protocols.io/view/high-quality-dna-extraction-protocol-from-recalcit-i8jchun
955 (Accessed January 3, 2025).

956 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear
957 R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
958 the Genomic Era. *Mol Biol Evol* **37**: 1530–1534.

959 Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, Pusey AE,
960 Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota with hominids.
961 *Science* **353**: 380–382.

962 Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes
963 using nanopore sequencing. *Nat Biotechnol* **38**: 701–707.

964 Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA,
965 Muzny DM, Gibbs RA, et al. 2017. The gut mycobiome of the Human Microbiome
966 Project healthy cohort. *Microbiome* **5**: 153.

967 Nieves-Ramírez ME, Partida-Rodríguez O, Laforest-Lapointe I, Reynolds LA, Brown EM,
968 Valdez-Salazar A, Morán-Silva P, Rojas-Velázquez L, Morien E, Parfrey LW, et al.
969 2018. Asymptomatic Intestinal Colonization with Protist Blastocystis Is Strongly
970 Associated with Distinct Microbiome Ecological Patterns. *mSystems* **3**.

971 Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie
972 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*
973 **33**: 290–295.

974 Piperni E, Nguyen LH, Manghi P, Kim H, Pasolli E, Andreu-Sánchez S, Arrè A, Bermingham
975 KM, Blanco-Míguez A, Manara S, et al. 2024. Intestinal Blastocystis is linked to
976 healthier diets and more favorable cardiometabolic outcomes in 56,989 individuals from
977 32 countries. *Cell* **187**: 4554-4570.e18.

978 Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for
979 reference-free profiling of polyploid genomes. *Nat Commun* **11**: 1432.

980 Saary P, Mitchell AL, Finn RD. 2020. Estimating the quality of eukaryotic genomes recovered
981 from metagenomic analysis with EukCC. *Genome Biol* **21**: 244.

982 Santin M, Figueiredo A, Molokin A, George NS, Köster PC, Dashti A, González-Barrio D,
983 Carmena D, Maloney JG. 2024. Division of Blastocystis ST10 into three new subtypes:
984 ST42-ST44. *J Eukaryot Microbiol* **71**: e12998.

985 Scanlan PD, Marchesi JR. 2008. Micro-eukaryotic diversity of the human distal gut microbiota:
986 qualitative assessment using culture-dependent and -independent analysis of faeces.
987 *ISME J* **2**: 1183–1193.

988 Scanlan PD, Stensvold CR. 2013. Blastocystis: getting to grips with our guileful guest. *Trends*
989 *Parasitol* **29**: 523–529.

990 Scanlan PD, Stensvold CR, Rajilić-Stojanović M, Heilig HGHJ, De Vos WM, O'Toole PW,
991 Cotter PD. 2014. The microbial eukaryote Blastocystis is a prevalent and diverse member
992 of the healthy human gut microbiota. *FEMS Microbiol Ecol* **90**: 326–330.

993 Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient
994 gene linkages support ctenophores as sister to other animals. *Nature* **618**: 110–117.

995 Šimková H, Câmara AS, Mascher M. 2024. Hi-C techniques: from genome assemblies to
996 transcription regulation. *J Exp Bot* **75**: 5357–5365.

997 Smit A, Hubley R, Green P. 2013. Repeatmasker Open-4.0. http://www.repeatmasker.org
998 (Accessed January 10, 2015).

999 Sommer F, Bäckhed F. 2013. The gut microbiota--masters of host development and physiology.
1000 *Nat Rev Microbiol* **11**: 227–238.

1001 Stensvold CR, Clark CG. 2020. Pre-empting Pandora's Box: Blastocystis Subtypes Revisited.
1002 *Trends Parasitol* **36**: 229–232.

1003 Stensvold CR, Suresh GK, Tan KSW, Thompson RCA, Traub RJ, Viscogliosi E, Yoshikawa H,
1004 Clark CG. 2007. Terminology for Blastocystis subtypes--a consensus. *Trends Parasitol*
1005 **23**: 93–96.

1006 Stevens CE, Hume ID. 1995. *Comparative physiology of the vertebrate digestive system*. 2nd ed.
1007 Cambridge University Press, Cambridge
1008 http://catdir.loc.gov/catdir/toc/cam026/95013969.html (Accessed October 16, 2023).

1009 Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of
1010 transposable element families, sequence models, and genome annotations. *Mob DNA* **12**:
1011 2.

1012 Tan KSW. 2008. New Insights on Classification, Identification, and Clinical Relevance of
1013 Blastocystis spp. *Clin Microbiol Rev* **21**: 639–665.

1014 Thorman AW, Adkins G, Conrey SC, Burrell AR, Yu Y, White B, Burke R, Haslam D, Payne
1015 DC, Staat MA, et al. 2023. Gut Microbiome Composition and Metabolic Capacity Differ
1016 by FUT2 Secretor Status in Exclusively Breastfed Infants. *Nutrients* **15**: 471.

1017 Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. 2022. DeepLoc 2.0:
1018 multi-label subcellular localization prediction using protein language models. *Nucleic*
1019 *Acids Res* **50**: W228–W234.

1020 Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from
1021 long uncorrected reads. *Genome Res* **27**: 737–746.

1022 Wang SE, Amir AS, Nguyen T, Poole AM, Simoes-Barbosa A. 2018. Spliceosomal introns in
1023 Trichomonas vaginalis revisited. *Parasit Vectors* **11**: 607.

1024 West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for
1025 eukaryotes from complex natural microbial communities. *Genome Res* **28**: 569–580.

1026 Wick RR, Holt KE. 2022. Polypolish: Short-read polishing of long-read bacterial genome
1027 assemblies. *PLOS Comput Biol* **18**: e1009802.

1028 Wilihoeft U, Campos-Góngora E, Touzni S, Bruchhaus I, Tannich E. 2001. Introns of
1029 Entamoeba histolytica and Entamoeba dispar. *Protist* **152**: 149–156.

1030 Yadav M, Verma MK, Chauhan NS. 2018. A review of metabolic potential of human gut
1031 microbiome in human nutrition. *Arch Microbiol* **200**: 203–217.

1032 Yang JK, Yoon HJ, Ahn HJ, Lee BI, Pedelacq J-D, Liong EC, Berendzen J, Laivenieks M,
1033 Vieille C, Zeikus GJ, et al. 2004. Crystal structure of beta-D-xylosidase from
1034 Thermoanaerobacterium saccharolyticum, a family 39 glycoside hydrolase. *J Mol Biol*
1035 **335**: 155–165.

1036 Yoshikawa H, Koyama Y, Tsuchiya E, Takami K. 2016. Blastocystis phylogeny among various
1037 isolates from humans to insects. *Parasitol Int* **65**: 750–759.

1038 Záhonová K, Low RS, Warren CJ, Cantoni D, Herman EK, Yiangou L, Ribeiro CA, Phanprasert
1039 Y, Brown IR, Rueckert S, et al. 2023. Evolutionary analysis of cellular reduction and
1040 anaerobicity in the hyper-prevalent gut microbe Blastocystis. *Curr Biol CB* **33**: 2449-
1041 2464.e8.

1042 Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018.
1043 dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic*
1044 *Acids Res* **46**: W95–W101.

1045 Zierdt CH. 1988. Blastocystis hominis, a long-misunderstood intestinal parasite. *Parasitol Today*
1046 *Pers Ed* **4**: 15–17.

1047 Zierdt CH, Williams RL. 1974. Blastocystis hominis: Axenic cultivation. *Exp Parasitol* **36**: 233–
1048 243.

1049 Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA
1050 genome assembler. *Bioinformatics* **29**: 2669–2677.

1051

1052