Hybrid assemblies of microbiome *Blastocystis* protists reveal evolutionary diversification reflecting host ecology

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

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

Abstract
The most prevalent microbial eukaryote in the human gut is *Blastocystis*, an obligate commensal protist also common in many other vertebrates. *Blastocystis* is descended from free-living stramenopile ancestors; how it has adapted to thrive within humans and a wide range of hosts is unclear. Here, we cultivated six *Blastocystis* strains spanning the diversity of the genus and generated highly contiguous, annotated genomes with long-read DNA-seq, Hi-C, and RNA-seq. Comparative genomics between these strains and two closely related stramenopiles with different lifestyles, the lizard gut symbiont *Proteromonas lacertae* and the free-living marine flagellate *Cafeteria burkhardae*, reveal the evolutionary history of the *Blastocystis* genus. We find substantial gene content variability between *Blastocystis* strains. *Blastocystis* isolated from an herbivorous tortoise has many plant carbohydrate metabolizing enzymes, some horizontally acquired from bacteria, likely reflecting fermentation within the host gut. In contrast, human-isolated *Blastocystis* have gained many heat shock proteins, and we find numerous subtype-specific expansions of host-interfacing genes, including cell adhesion and cell surface glycan genes. In addition, we observe that human-isolated *Blastocystis* have substantial changes in gene structure, including shortened introns and intergenic regions, as well as genes lacking canonical termination codons. Finally, our data indicate that the common ancestor of *Blastocystis* lost nearly all ancestral genes for heterokont flagella morphology, including cilia proteins, microtubule motor proteins, and ion channel proteins. Together, these findings underscore the huge functional variability within the *Blastocystis* genus and provide candidate genes for the adaptations these lineages have undergone to thrive in the gut microbiomes of diverse vertebrates.

Introduction

The vertebrate gut is host to a diverse ecosystem of bacteria, archaea, viruses, and microbial eukaryotes. Compared to bacteria, the biology and impact of microbial eukaryotes in the human microbiome is underexplored. The vertebrate gut is host to a diverse microbial community that has broad impacts on the physiology and development of the host. For example, gut microbes can metabolize compounds that are not digestible by the host, making

previously inaccessible nutrients available[1]. Furthermore, gut microbes and their metabolic byproducts have profound impacts on the host, both through interactions with the immune systems and from downstream effects of circulating microbial metabolites[2]. The composition of the vertebrate gut microbiome therefore impacts host phenotype, and is shaped by host diet, host physiology, and the host immune system[3–5]. Indeed, microbiome composition is highly variable across animals as a function of phylogeny, physiology, diet, and behavior.

The microbial communities in vertebrate guts contain bacteria, archaea, viruses, and microbial eukaryotes. 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 protists across vertebrates[6]. Eukaryotic protist members of the microbiome have been characterized, including tritrichomonads with direct immune system interactions in mammals[7], metamonads which digest lignocellulose in termites[8], and ciliates which ferment carbohydrates in11/21/2023 12:26:00 AM ruminants[9].

The protist *Blastocystis* is the most prevalent eukaryotic microorganism in the human gut, as demonstrated by eukaryotic marker gene surveys and shotgun metagenomic analyses[10]. Prevalence estimates within industrialized communities range from 10-50%, and are higher within populations living in less industrialized settings[11–13]. In addition to humans, *Blastocystis* protists colonize a large range of animals including other mammals, birds, reptiles, fish, amphibians, and insects[14–16]. *Blastocystis* is a genetically diverse genus that has successfully colonized guts from most vertebrate lineages, spanning a huge array of gut physiologies, host enteric immune systems, host body temperatures, and host diets[17]. As obligate commensals, *Blastocystis* do not grow outside of the host, though they form cysts that can persist in the environment. This widespread colonization also requires *Blastocystis* to contend with diverse microbial communities within different vertebrate guts. Nonetheless, very little is understood about how *Blastocystis* contributes to the gut microbiome environment and interacts with other gut microbiota.

Blastocystis and their close relatives the slopalinids comprise the Opalinata lineage of stramenopile protists. Like *Blastocystis,* slopalinids are obligate animal gut commensals, though slopalinids are predominantly found in amphibians and lizards[18]. While the majority of stramenopiles have characteristic flagella and surface hairs, *Blastocystis* is non-motile and lacks flagella or surface hairs, and is predominantly observed in culture as round cells with a large central vacuole superficially resembling a yeast cell[19]. This appearance contrasts with the morphological diversity of close relatives within the Opalinata, most of which are motile and flagellated.

To understand how the *Blastocystis* genus has evolved its unique morphology and adapted to a diversity of hosts, we cultivated 6 *Blastocystis* strains, five isolated from humans and one isolated from the herbivorous tortoise *Testudo horsfeldii.* We generated highly contiguous, annotated genomes using a combination of long-read DNA-seq, Hi-C, and RNA-seq. These highly contiguous and complete genomes allow us to investigate patterns of genome evolution within and between subtypes. These isolates span the diversity of the *Blastocystis* genus, enabling us to investigate *Blastocystis* evolution and determine how *Blastocystis* has adapted to thrive in animal hosts with different diets, physiologies, and resident microbiota.

Materials and methods

Strain cultivation

Blastocystis strains NandII (ATCC 50177), JDR (ATCC 50587), BT1 (ATCC 50608), DL (ATCC 50626), and NMH (ATCC 50754) were obtained from the American Type Culture Collection. The JDR strain was labeled as subtype 3 based on rRNA analysis, but 18S sequencing and subsequent genome sequencing revealed in fact that this strain belongs to subtype 1. All strains were isolated from human sources. Cultures were incubated under anaerobic conditions at 37C on biphasic egg slant media overlaid with Locke's solution supplemented with 20% heat inactivated horse serum. ST1 and ST4 strains were maintained in axenic culture, while ST3 and the tortoise isolate were grown in xenic culture alongside other microbes present in the original samples. These xenic culture strains proved resistant to axenization with antibiotics and may require other microbes for growth.

The tortoise *Blastocystis* isolate was obtained from fecal material from a domestic Russian tortoise that also contained *Trichomitus batrachorum.* Protists were isolated from tortoise stool by resuspending a fresh stool sample in sterile PBS. Large particulate was then removed using a 70µm cell strainer, and the strained sample was then centrifuged at 700xg for 5 minutes at room temperature. The pellet was then washed 3 times with PBS before being resuspended in 80% Percoll. The percoll-resuspended sample was centrifuged at 1000xg for 10 minutes at room temperature. Protists were then removed from the top of the percoll layer, washed in PBS, and resuspended in growth medium.

DNA extraction and sequencing

High molecular weight DNA was isolated from *Blastocystis* cultures with a modified CTAB-based protocol (adapted from https://dx.doi.org/10.17504/protocols.io.i8jchun) that minimizes polysaccharide contamination. *Blastocystis* cells were collected from cultures by pelleting with centrifugation at 800*g* for 5 minutes and snap frozen at -80˚C. Frozen pellets were gently resuspended in two volumes of a high salt CTAB lysis buffer (100 mM Tris pH 8, 1.2 M NaCl, 10 mM EDTA, 4% CTAB, 1% PVP-40). 200 µg/mL RNAse A and 200 µg/mL proteinase K were added before incubation at 60˚C for 90 minutes. One volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added and samples were mixed end over end for 15 minutes. Samples were centrifuged at 17000*g* for 10 minutes and the upper aqueous layer was recovered. Two additional extractions were performed with pure chloroform, recovering the upper aqueous layer each time. DNA was next precipitated with 2 volumes of a low salt CTAB precipitation buffer (100 mM Tris pH 8, 10 mM EDTA, 2% CTAB) by end over end rotation at 60˚C for 2.5 hours. DNA was pelleted by centrifugation at 17000*g* for 20 minutes. DNA pellets were washed twice with 75% ethanol and resuspended in 10 mM Tris pH 8. Genomic DNA was further cleaned with 1.8 volumes of Ampure XP beads according to the manufacturer's protocol (Beckman Coulter). RNA was isolated from *Blastocystis* cultures with a Trizol extraction following manufacturer's protocols (Invitrogen).

Nanopore sequencing libraires were prepared using the Genomic DNA by Ligation protocol (Oxford Nanopore Technologies SQK-LSK109) following the manufacturer's instructions. Each library was sequenced on a full FLO-MIN106D R9 flow cell on a MinION sequencer. Base calling was performed using guppy version 5.0.16. Short read Illumina DNA and RNA libraries were prepared using the Roche KAPA HyperPlus PCR-free and KAPA mRNA capture kits, respectively, following manufacturer's instructions. RNA integrity and library fragment size

distribution were quantified using an Agilent 2100 Bioanalyzer. Illumina DNA and RNA libraries for all human *Blastocystis* isolates were sequenced on an Illumina NextSeq 500, and libraries for the *Blastocystis* tortoise isolate were sequenced on an Illumina MiSeq v3.

Genome assembly and decontamination

Basecalled nanopore reads were assembled using Flye version 2.9.2, with the --meta option enabled for xenic cultures (*Blastocystis* DL, NMH, and the tortoise *Blastocystis* isolate)[20,21]. Assembly polishing was performed using four rounds of nanopore-read polishing using racon version 1.4.3[22], three rounds of Illumina DNA polishing using PolyPolish version 0.5.0[23], and 1 round of short-read polishing using POLCA as implemented in MaSuRCA version 4.0.9[24].

For the xenic assemblies for the DL, NMH, and tortoise isolates, we took a conservative approach to identify and remove non-*Blastocystis* sequences, combining multiple lines of evidence. We used three different approaches to determine the taxonomic origin of each contig in the assembly: MetaEuk protein prediction and taxonomic assignment[25], GTDB-Tk classification of each contig[26], and querying each contig against the NCBI nt database using BLAST. GTDB-Tk classification of contigs is listed in Table S1. In addition, we used the *k*-mer based classifier EukRep[27] to determine contigs that were likely bacterial in origin, though mitochondrial contigs filtered by EukRep were retained in the assembly. Finally, we leveraged the poly-A purification strategy of our RNA-seq libraries and aligned RNA-seq reads to assemblies in order to exclude contigs with no aligned reads.

In addition to bacteria, the xenic tortoise *Blastocystis* culture contained a low levels of the tritrichomonad *Trichomitus batrachorum*. To remove contigs with *T. batrachorum* sequences, we combined the taxonomic approach described above with a metagenomics binning approach using MetaBat2[28] and by aligning the *T. batrachorum* transcriptome[29] to the assembled contigs. One metagenomic bin identified by MetaBat2 had strong agreement with a eukaryotic origin according to EukRep and a stramenopile according to MetaEuk taxonomy, but did not have any aligned *Trichomitus batrachorum* transcripts. All contigs outside of this bin were excluded from the final assembly.

Gene annotation

Blastocystis genes occasionally do not encode stop codons, which confounds prediction via open reading frame and requires RNA-seq based evidence[30,31]. Without consideration of lack of stop codons, the gene annotation errors that are common are insertion of erroneous intron and combining adjacent genes. To accurately annotate the *Blastocystis* genomes, we used a combination of *de novo* gene prediction with Braker3 and transcriptome assembly, combining all sets of gene models with Mikado[32,33].

RNA-seq reads were aligned to assemblies with STAR[34], with a minimum of 3 supporting reads required for each junction. Transcript assemblies were generated with Stringtie[35], 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 at regions of zero RNA-seq coverage. Assembled transcripts less than 100 bp long were discarded. To generate *de novo* gene prediction, the Braker3 pipeline was run with OrthoDB11 Eukaryota proteins as protein

evidence and the aligned RNA-seq as transcript evidence[33,36]. The Mikado scoring file (available in code repository) was optimized to choose transcripts with strictly RNA-seq supported introns, but to prioritize genes containing stop codons as long as the intron boundaries were strictly supported.

Functional annotatation and orthology inference

Proteins from *Cafeteria burkhardae* BVI (formerly *Cafeteria roenbergensis* BVI), and predicted proteins from *Proteromonas lacertae* were used for comparative genomics analyses[37,38]. The genome assembly, transcripts, and proteins for *C. burkhardae* was downloaded from GenBank under accession GCA_008330645.1. The genome sequence for *Proteromonas lacertae* was downloaded from GenBank under accession GCA_002245135.1. As no predicted proteins or RNA-seq data was available for *P. lacertae*, genes were predicted from this genome using the Braker3 pipeline, using OrthoDB11 Eukaryota protein sequences as evidence. Proteins from all *Blastocystis* assemblies, *C. burkhardae*, and *P. lacertae* were annotated using a combination of InterProScan version 5.62[39], the CDD suite of tools[40], DeepLoc2 to predict protein localization[41], and dbCAN version 4.0.0 to annotate carbohydrate active enzymes[42]. Orthologous sequences were determined using OrthoFinder2.

Strain phylogeny

The *Blastocystis* genus phylogeny was constructed using 18S rRNA sequences deposited in GenBank (Table S2) along with 18S rRNA sequences extracted from each *Blastocystis* assembly and from the *Proteromonas lacertae* genome as an outgroup. Sequences were aligned with MAFFT[43] and a maximum likelihood phylogeny constructed using IQ-TREE version 2.1.4[44] using ModelFinder[45] to determine the best-fit substitution model, 1,000 ultrafast bootstrap replicates, and rooted on *P. lacertae*. Branches with less than 80% bootstrap support were collapsed.

Horizontal gene transfer screens and gene tree reconstruction

Blastocystis genomes were screened for bacterial contamination and for possible horizontal gene transfer events using a modified version of an alien index score[46,47]. The alien index score compares the best bit score of a query sequence to a group of related taxa versus to all other unrelated groups, and normalizes this score by the maximum possible bit score of the query sequence. It is determined by the formula $AI = \frac{bbsO - bbsI}{bbsS}$, where bbsO is the bit score of the best hit to a species outside of the group lineage, bbsI is the bit score of the best hit to a species within the group lineage (skipping self hits), and bbsS is the bit score of the query aligned to itself. The alien index was calculated for each *Blastocystis* protein by querying proteins against the NCBI nr database using DIAMOND version 2.1.6 [48](code availability: https://github.com/allind/alienindex). The self-group was designated as *Blastocystis* (taxid: 12967), and the group lineage was set as all eukaryotes (taxid: 2759).

Phylogenetic trees for putative horizontal gene transfer candidates were constructed from all blastp hits using the NCBI NR database clustered at 90% sequence identity and 90% sequence length. Sequence hits that were outliers in length were excluded from downstream analyses. Protein sequences were aligned with MAFFT, and maximum likelihood phylogenies

were constructed using IQ-TREE version 2.1.4 using ModelFinder to determine the best-fit substitution model and 10,000 ultrafast bootstrap replicates. Branches with bootstrap support less than 80% were collapsed.

Results

Capturing Blastocystis diversity

In order to investigate the evolution of the *Blastocystis* genus, including its initial divergence as well as its later diversification, we acquired six divergent *Blastocystis* strains for sequencing. Strains were cultured anaerobically on egg media overlayed with a balanced salt solution and 25% heat-inactivated horse serum. Five of these strains correspond to 3 different previously defined "subtypes" (genetically similar groups historically identified in mammalian avian species[49]). These five strains were originally cultured from human fecal samples and were obtained from the American Type Culture Collection. The sixth strain was obtained from a domestic Russian tortoise *Testudo horsfeldii*, and 18S sequencing determined it was a member of a group distantly related to the human isolates that does not belong to a described mammalian or avian subtype (Figure 1A). Two human isolated samples belong to subtype 1 (ST1), two belong to subtype 3 (ST3), and one belongs to subtype 4 (ST4), together representing 3 of the 4 most commonly detected *Blastocystis* subtypes in humans. Analysis of these cultures by light microscopy revealed predominantly vacuolar *Blastocystis* cells, containing a large central vacuole and a thin peripheral ring of cytoplasm. While *Blastocystis* have been observed to adopt a variety of morphologies, vacuolar morphologies dominate cultured samples. The tortoise isolate appeared morphologically similar to the human isolates, though cells in this culture were always found in multi-cell clumps, unlike the human isolates (Figure 1B).

Contiguous and complete Blastocystis genomes

Table 1. Genome assembly statistics for *Blastocystis* isolates.

Strain	Subtype	Isolate source	Scaffold counts	Assembly size (Mb)	Assembly N50 (Kb)	GC content (% GC)	BUSCO completeness	Protein-coding genes
JDR	1	Human	114	17.5	550	54.5%	61.20%	7891
NandII	1	Human	54	16	939	54.7%	61.20%	7157
NMH	3	Human	36	14.6	582	51.3%	58.00%	6571
DL	3	Human	56	14	506	51.6%	55.70%	6154
BT1	4	Human	34	15.3	882	39.4%	57.60%	7421
Blastoise	N/A	Russian tortoise	33	33.3	1323	23.1%	43.20%	8547

We prepared high quality, high molecular weight DNA, as well as RNA, from the six diverse *Blastocystis* cultures. DNA was sequenced with Nanopore long-read technology. Short-read DNA and RNA sequencing was also performed for assembly polishing and gene annotation. Hi-C data was additionally generated to help resolve scaffolds for one strain. For the three strains from xenic cultures observed to contain other microbes, sequencing and assembly

revealed the presence of specific bacteria, archaea, and viruses (Table S1). Non-eukaryotic sequence was detected via multiple methods and excluded from final assemblies (see Methods). This combination of sequencing technologies enabled us to generate highly contiguous genomes for all six subtypes. The N50 of these assemblies ranged from 506 Kb to 1.3 Mb, achieving a much higher assembly quality than currently available *Blastocystis* genome datasets in NCBI (median: 37 Kb) (Table 1). The median number of scaffolds was 45, compared to a median of 580 for previously sequenced genomes. Interestingly, the sequenced *Blastocystis* strains showed substantial variability in genome size, from 14.0 to 32.2 Mb. Strains in the same subtype exhibited similar genome sizes: ST3 isolates were 14.0 and 15.3 Mb, and ST1 isolates were 17.2 and 16.0 Mb in size. The tortoise isolate was larger than any human-derived isolate at 32.2 Mb. BUSCO completeness estimates ranged from 43% to 61%, similar to previous analyses of *Blastocystis* and related species (Table 1)[31,37]. The BUSCO estimate for the larger-sized tortoise isolate, was comparable to the human isolates, indicating the difference in genome size is not due to increased genome completeness.

Previous genome sequencing studies have raised questions about the ploidy of *Blastocystis* and found evidence of segmentally duplicated blocks of genes. With the highly contiguous genomes and multiple isolates of individual subtypes here, we investigated questions of ploidy and conservation of synteny both within *Blastocystis* subtypes and across the genus. *K*-mer frequency-based analysis using GenomeScope[50] demonstrated a strong diploid signal in all *Blastocystis* isolates, with limited evidence for aneuploidy or polyploidy. We found that strains within ST3 and within ST1 were highly syntenic (Figure S1). However, synteny between subtypes was much lower.

We used a hybrid method of de novo gene prediction and RNA transcript evidence to annotate *Blastocystis* genes (see Methods). Because a substantial fraction of genes in *Blastocystis* subtypes 1 and 7 do not have DNA-encoded stop codons[30,31] (discussed below), open reading frame based gene predictions from DNA sequence are inaccurate. Leveraging our RNA sequencing data, we were able to generate a comprehensive gene annotation for each *Blastocystis* genome (see Methods). The number of genes per genome ranged from 6,154 to 8,547. There was a median of 5-7 introns per gene across isolates, and the median protein length ranged from 352 amino acids to 438 amino acids across isolates.

To determine the level of sequence divergence of the isolates here, we calculated average sequence identity over 1,350 single-copy orthologous genes identified using OrthoFinder2[51] that are present in all *Blastocystis* isolates. Isolates of the same subtype were highly similar; the average sequence identity between isolates within ST1 and ST3 was 99%. Between subtype amino acid identity dropped to 71% between subtypes 3 and 4, to 63% between ST1 and both ST3 and ST4, and then to 42-43% between the tortoise isolate and all human isolates (Figure S2).

The evolution of human-hosted Blastocystis

We used the 18S rRNA gene extracted from sequenced genomes as a phylogenetic marker to construct a phylogeny of the *Blastocystis* genus from 63 *Blastocystis* rRNA sequences deposited in GenBank as well as the 18S rRNA sequence from the genomes here (Figure 1A). The 18S rRNA sequence of our tortoise *Blastocystis* strain was substantially diverged from the human subtype sequences and is not closely related to any designated mammal or avian

subtypes. This divergence is reflected in multiple different aspects of the *Blastocystis* genomes, including the GC content, which varies substantially between *Blastocystis* isolates. The GC content of human isolates ranged from 39-55%, while the tortoise isolate had a substantially lower GC content of 23% (Table 1). The GC content of the tortoise isolate is substantially lower than the genomes of most sequenced stramenopiles; of the 299 stramenopiles with sequenced genomes deposited in GenBank, only three genomes have lower GC content.

To explore how the *Blastocystis* genus has emerged and diversified, we compared the *Blastocystis* genomes generated in this work to those of two other sequenced stramenopiles: *Cafeteria burkhardae*, a free-living heterotrophic marine flagellate[38], and the recently sequenced *Proteromonas lacertae*, a reptile gut commensal closely related to *Blastocystis*[37]. These organisms are the closest relatives to *Blastocystis* that have sequenced genomes. Comparison between these genomes allowed us to reconstruct three evolutionary transitions that have occurred during stramenopile evolution: the transition from a free-living aquatic lifestyle to an anaerobic lifestyle inside the animal gut (comparing Cafeteria to *Blastocystis*/Proteromonas), the change in cell morphology from heterokont flagellate to aflagellate, spherical cells (comparing Cafeteria/Proteromonas to *Blastocystis*), and adaptations to different animal hosts (comparing *Blastocystis* tortoise to *Blastocystis* human) (Figure 1B). We find that each of these transitions is accompanied by changes to numerous gene families (Table S3).

Changes to gene density and structure within Blastocystis

In *Blastocystis* ST1 and ST7, it has been reported that a fraction of protein coding genes (15% for ST7 and 26% for ST1) lack canonical DNA-encoded termination codons[30,31]. A proposed mechanism for translation of these genes is that the translational machinery instead uses a U or a UG sequence in mRNA, along with one or two adenines from the poly-A tail, to generate UAA and UGA on-the-fly stop codons[30]. Accompanying this loss of canonical stops is a highly conserved motif roughly 5 base pairs downstream of transcription stop sites of most expressed sequence tags (ESTs) in *Blastocystis* ST7 and most genes in *Blastocystis* ST1[30,31]. The high conservation of this motif suggests that the motif may be directing 3' mRNA cleavage and polyadenylation factors to the pre-mRNA to direct stop codon incorporation. Loss of termination codons in these genes likely reflects that the selection pressure maintaining them has been lost.

To investigate whether this mechanism is common across *Blastocystis* strains or unique to certain subtypes, we investigated all six sequenced strains for the downstream sequence element. Subtypes 1, 3, and 4 had a highly conserved element immediately downstream of 55-71% of all protein coding genes (Figure 2A-B; Figure S3). RNA-seq supported stop codon-less gene models for a fraction (15%-37%) of genes. This is in line with previous estimates that have suggested 15% for ST7 and 26% for ST1[30,31]. Gene ontology enrichment analyses found no enrichment for genes lacking stop codons versus genes containing stop codons, suggesting there is no relationship between stop codon loss and gene function. Unexpectedly, we found no such DNA motifs or evidence of transcripts lacking stop codons in the tortoise *Blastocystis* isolate. This finding suggests the unique transcriptional cleavage machinery previously hypothesized to exist in all *Blastocystis* evolved after the divergence of tortoise and human isolate lineages.

Another genomic feature that distinguishes the tortoise isolate and human isolate divergence is genome size and GC content. The GC content of the tortoise isolate is much lower than the human isolates and generally for stramenopiles, and is similar to the lizard commensal *Proteromonas lacertae*. The precise cause of GC content variability among microbes has been subject to debate, with multiple proposed mechanisms including optimal growth temperature[52,53]. The lower GC content of these reptile associated microbes relative to human associated microbes may reflect differences in host body temperature, though other possibilities cannot be excluded. The genome of the tortoise isolate is twice the size of the human isolates, but with a similar number of predicted genes. We therefore investigated what genomic elements were the source of this larger genome. We found that the almost all introns in human ST1, ST3, and ST4 *Blastocystis* are exactly 30 bp long, in agreement with previous estimates of intron size in in *Blastocystis* ST1 and ST7[31,54]. However, introns in the tortoise *Blastocystis* isolate are longer and more variable in length (Figure 2C). To directly compare introns from the same genes, we examined 1,350 single-copy orthologous genes conserved in all *Blastocystis* strains sequenced here. The tortoise *Blastocystis* isolate genes had a median of 7 introns, while human *Blastocystis* isolate genes had a median of 5 introns (Figure S4). Across these genes, each human isolate has a median intron size of 30 bp, with very little deviation. However, the tortoise *Blastocystis* isolate has a median intron length of 99 bp.

In addition to longer introns, we found the intergenic spaces between genes are much longer in the tortoise *Blastocystis* isolate than in human isolates (Figure 2D). The median intergenic length in the tortoise isolate is 700 bp, while the median intergenic length in the human isolates ranges from 137-190 bp. Intergenic lengths in the human isolates can frequently be extremely small, only spanning tens of base pairs (Figure 2D; Figure S3). Together, the total amount of DNA spanned by introns and intergenic regions is 16.5 Mb in the tortoise *Blastocystis* isolate, whereas these regions span 3.9-5.9 Mb in human *Blastocystis* isolates. Considered together, the larger introns and intergenic distance in the tortoise isolate explains the majority of the difference in the genome size of this isolate relative to the human isolate.

Subtype-specific genes reveal specialization within the Blastocystis genus

In addition to changes in genome structure, we find differences in gene content between *Blastocystis* isolates that reflect their ecological niche. One of these differences is in strains' enzyme complement to digest plant carbohydrates present in different host diets. Compared to human isolates, the tortoise *Blastocystis* isolate contains a larger number and variety of genes dedicated to breaking down plant-specific carbohydrates. Specifically, this genome contains 6 copies of polygalacturonase, a pectinase that breaks down carbohydrates in plant cell walls (Figure 3), while human-isolated strains have no copies of this gene. All of these genes are predicted to encode secreted enzymes. The polygalacturonase enzyme was originally characterized in plants, but is also present in phytopathogenic microbes and in bacterial members of ruminant microbiomes[55,56]. While the origin of this gene family in the tortoise *Blastocystis* genome is unclear, these genes are homologous to polygalacturonases in other stramenopiles and also in plants, suggesting they did not arise through horizontal gene transfer. In addition to polygalacturonase genes, the tortoise *Blastocystis* genome contains 7 copies of beta-D-xylosidase genes, which are required for degrading xylan, a carbohydrate specific to plant hemicellulose[57]. This gene is present in fewer copies in all sequenced human *Blastocystis*

isolates; both isolates of ST1 and the subtype 4 isolate contain 1 copy each, while both ST3 isolates contain no beta-xylosidase genes (Figure 3). Together, the presence of plant-specific carbohydrate digesting enzymes in an isolate from an herbivorous reptile strongly suggests that, like the bacterial microbiota, *Blastocystis* utilize dietary plant fiber as a carbon source and participate in the hindgut plant fiber fermentation that is characteristic of herbivorous tortoises[58].

In addition to changes that reflect differences in host diet, host physiology has likely impacted the evolution of *Blastocystis*. Human-isolated *Blastocystis* strains contain many more copies of protein folding chaperones, specifically in the heat shock protein 70 (Hsp70) category, than the tortoise isolate or other stramenopiles (Figure 3). The tortoise isolate contains 15 genes annotated with the Hsp70 InterPro domain (IPR013126), while ST1 isolates contain 72-76 Hsp70 genes, ST3 isolates contain 49-56 copies, and ST4 contains 73 copies. *C. burkhardae* contains 11 Hsp70-annotated genes. These differences between human-isolated *Blastocystis* and tortoise-isolated *Blastocystis* correlate with differences in host digestive tracts: endothermic, higher temperature human versus ectothermic, lower temperature tortoise (Figure 1). As Hsp70 genes increase in expression during heat stress in *Blastocystis*[59] and other parasites such as *Plasmodium falciparum*[60], we hypothesize that the expansion of this family of chaperone genes has adapted human-colonizing *Blastocystis* subtypes for survival at mammalian body temperatures.

Ongoing horizontal gene transfer from bacteria in the Blastocystis genus

Horizontal gene transfer from bacteria has played an important role in *Blastocystis* evolution, particularly with respect to survival in an anaerobic environment[61]. With high-quality genomes of *Blastocystis* isolates that span the genus, we can estimate times of these transfer events.

Previous work in *Blastocystis* ST1 strain NandII identified 74 genes that have undergone horizontal gene transfer in this organism. Of these genes, 65 had unambiguous reciprocal best BLAST hits in our long-read *Blastocystis* ST1 strain NandII assembly (Figure 4A). Of these 65 gene transfers, 3 are present in all *Blastocystis* subtypes, *P. lacertae*, and *C. burkhardae*, indicating these transfer events happened prior to the divergence of these species. Nineteen transfers are present in *Proteromonas lacertae* and in all *Blastocystis* subtypes sequenced, indicating that these transfers occurred early in the evolution of Opalinata, which coincided with a lifestyle transition from free-living aerobic to obligately host-associated and anaerobic. Twelve of the horizontally transferred genes are not present in *P. lacertae* or *C. burkhardae*, but are present in all *Blastocystis* subtypes, indicating that these transfer events occurred at the emergence of the *Blastocystis* genus. Eight of the horizontally transferred genes were transferred into *Blastocystis* after the divergence of the tortoise *Blastocystis* isolate and the human isolates; these 8 genes are not present in *C. burkhardae*, *P. lacertae*, or the tortoise *Blastocystis*. The remaining 23 genes show more complex patterns of gain and loss, including one gene (an S-adenosyl-L-methionine-dependent methyltransferase) present in both isolates of ST1 and no other isolates, suggesting this was recently acquired by this subtype. Thus, gene transfer has been ongoing throughout Blastocystis evolution.

We also identified six instances of horizontal gene transfer from bacteria into the branch leading to the tortoise *Blastocystis* lineage (Figure 4A). We identified horizontally transferred

genes with a conservative approach including genomic context to rule out contamination (see Methods), so this number may be an underestimate. Four of these transfers include glycoside hydrolase genes, while the remaining two transfers are genes that encode a formate-nitrate transporter and a protein containing a conserved domain of unknown function (Figure 4B-C, Figure S5). The four glycoside hydrolase gene families that were transferred from bacteria into the lineage leading to the tortoise isolate are three beta-D-xylosidase genes, two alpha-L-rhamnosidase genes, one glucuronyl hydrolase gene in the GH88 CAZY family, and one gene belonging to the recently characterized GH154 CAZY family.

As discussed previously, the tortoise isolate contains 7 copies of the cellulolytic plant-fiber degrading beta-D-xylosidase enzyme. We find that three of these beta-D-xylosidase genes do not have eukaryotic homologs, and are bacterial in origin (Figure S5A). The bacterial donor is ambiguous, but the majority of the bacterial genes grouping with the *Blastocystis* gene are from the Bacteroidales order, which are prevalent in the animal gut. These genes have an additional eukaryotic homologue within sponges, which groups together with the *Blastocystis* genes within the gene tree. This gene may have been transferred multiple independent times from bacteria to eukaryotes and this grouping is an artifact of the long eukaryotic branches in the gene tree, or alternatively transfer may have occurred between the eukaryotes.

The tortoise *Blastocystis* genome is the only sequenced *Blastocystis* genome to contain several classes of glycoside hydrolases, including alpha-L-rhamnosidase genes, GH88 glucoronyl hydrolase genes, or GH154 glycoside hydrolase genes. The alpha-L-rhamnosidase gene family functions to cleave alpha-L-rhamnose from a large diversity of molecules and is present across the tree of life. The tortoise *Blastocystis* genome contains two of these genes that are not homologous to any eukaryotic genes. These genes group together with sequences from Bacillota (syn. Firmicutes), and the likely bacterial donor to *Blastocystis* was a bacteria in the Bacillota phyla (Figure 4B, Figure S5B). The GH88 glucoronyl hydrolases gene present in the tortoise *Blastocystis* isolate is not homologous to any eukaryotic genes. The bacterial donor for this gene is not clear, as the *Blastocystis* gene groups together with multiple anoxic or anaerobic environmental and gut bacteria from different lineages in the gene tree (Figure S5). The recently described GH154 glycoside hydrolase gene family, characterized in *Bacteroides thetaiotamicron* as acting on plant-derived arabinogalactan glycans[62], is also present exclusively in the tortoise *Blastocystis* isolate. The G154 gene is not homologous to any eukaryotic genes and was likely transferred into the tortoise *Blastocystis* lineage from bacteria in the Bacteroidota (syn. Bacteroidetes) phylum (Figure 4C, Figure S5D).

In addition to glycoside hydrolases acquired via horizontal gene transfer, the tortoise *Blastocystis* genome contains a formate-nitrate transporter and a gene containing the domain of unknown function DUF711. The formate-nitrate transporter gene has 3 eukaryotic homologues, including genes in two species of Anaeromoeba, a free-living anaerobic relative of parasitic metamonads[63], and one gene in the green algae Ostreobium quekettii (Figure S5E). The bacterial donor of this gene into *Blastocystis* is unclear. The presence of this gene in other eukaryotes suggests that this gene has been transferred multiple independent times. The DUF711-containing gene does not have any eukaryotic homologues, and the bacterial donor is unclear (Figure S5F). Together, these results indicate that the functions of gene transfers into Blastocystis differ systematically across the phylogeny.

Expansion of cell-surface and glycosylation gene families within Blastocystis

Several gene families that have expanded across the *Blastocystis* genus and are present in high copy numbers encode proteins that are either present on the cell surface or glycosylate proteins on the cell surface.

One gene family that is present in high copy numbers in *Blastocystis* but entirely absent from the genomes of its close relatives is the GT49 CAZyme glycosyltransferases. These are related to the metazoan glycosyltransferase LARGE, known best for glycosylating the extracellular glycoprotein dystroglycan[64]. This gene family is absent in *Cafeteria burkhardae* and *Proteromonas lacertae* but is present at high copy numbers in *Blastocystis*, with 20 copies in the tortoise isolate, 36-42 in ST3, 40-60 in ST1, and 66 in subtype 4 (Figure 5). While this gene family is not present in other closely related stramenopiles, it is homologous to GT49 glycosyltransferases found in other eukaryotic protists and is unlikely to have originated from horizontal gene transfer.

In addition to glycosyltransferases, *Blastocystis* and *Proteromonas lacertae* have high copy numbers of cadherin-like genes with a characteristic transmembrane domain and repetitive immunoglobulin domains. This gene family is present in high but variable copies in all *Blastocystis* subtypes from 31-41 in ST1, 38-41 in ST3, 117 in ST4, and 91 in the tortoise isolate (Figure 5). This gene family is present in 56 copies in *P. lacertae*. Some genes in this gene family also contain galactose-binding domains at the C-terminal end of the protein. Genes in this family have no eukaryotic homologues outside of *Blastocystis* or *Proteromonas*, and there are distant homologues within bacteria. Thus, the family may either be unique to this group of organisms or have distant origins within bacteria.

Finally, two families of secreted cysteine proteases, C13 legumain-like and C1A papain-like, are present in high copy numbers in all *Blastocystis* isolates and in *P. lacertae*, both of which inhabit animal guts, but are absent from the free-living aquatic *C. burkhardae*. *Blastocystis* strains and P. *lacertae* each have between 22-39 copies of C1A cysteine proteases and between 17-30 copies of C13 legumain-like cysteine proteases (Figure 5). Collectively, these results underscore the importance of cell surface proteins in stramenophile lineages inhabiting the gut of their vertebrate hosts.

Loss of morphological genes in Blastocystis

Blastocystis has a unique morphology for a stramenopile, as it lacks stereotypical motile flagella and tubular hairs, and instead is spherical and non-motile. This morphology is reflected by massive gene loss in this organism; in our sequenced genomes, we find all *Blastocystis* isolates sequenced have experienced a dramatic loss of cilia and flagella related genes, as well as a substantial reduction of both dynein and kinesin microtubule motor proteins, in agreement with recent analyses[37] (Figure 6). These losses correlate with the morphological changes from stereotypical stramenopiles. This loss is present across all *Blastocystis* isolates and is not variable, demonstrating that the entire *Blastocystis* genus experienced almost complete loss of these genes, retaining what is likely the minimal amount of motor proteins required for intracellular transport.

In addition to loss of flagella and cilia, we observed that *Blastocystis* has experienced a significant reduction in the number of voltage-gated ion channels relative to both *C. burkhardae* and *P. lacertae*. The number of genes annotated with the GO term GO:0006811 - monoatomic

ion transport is 54 in *C. burkhardae*, 43 in *P. lacertae*, but ranges from 4-7 in all *Blastocystis* species. While the majority of the ion transport proteins are predicted to be cell-membrane targeted in both *C. burkhardae* and *P. lacertae*, almost all remaining ion transport proteins in *Blastocystis* are predicted to be localized to either mitochondrial or vacuolar membranes. As protists rely on voltage-gated ion channels for cilial movement[65], we hypothesize that this loss of ion channel proteins in *Blastocystis* is linked to the loss of cilia and flagella.

Discussion

The stramenopile protist *Blastocystis* is the most prevalent protist in the human gut and is widespread throughout vertebrates. Here we present contiguous and complete genomes of 6 *Blastocystis* strains, 5 isolated from humans and one from an herbivorous tortoise. Comparative genomic analyses with other closely related stramenopiles revealed evolutionary changes that the *Blastocystis* genus has undergone to colonize a wide variety of animal guts.

The herbivorous tortoise isolate of *Blastocystis* diverged substantially from the human isolated *Blastocystis,* and has a number of genomic features not seen in the human isolates. We found substantial changes in genome size, with human isolates possessing significantly shorter introns and intergenic regions than the tortoise isolate. It is not clear whether introns and intergenic regions have shrunk in size after the split of the tortoise/human isolates, or if they have lengthened along the tortoise lineage. Genomic streamlining is a frequently observed process in symbiotic and parasitic organisms[66,67], and some parasitic eukaryotes have entirely or almost entirely lost introns[68,69]. While the free-living *C. burkhardae* has both longer introns and intergenic regions than any *Blastocystis* isolate, the lizard gut symbiont *P. lacertae* has introns and intergenic regions that are a similar length to the human *Blastocystis* isolates. The phenomenon of intergenic and intron size shrinking in *Blastocystis* likely reflects this general process, but the unexpected differences between different host-associated *Blastocystis* isolates and from host-associated *P. lacertae* highlights that this is an ongoing evolutionary process that can proceed at different rates. The tortoise isolate represents a middle ground between the free-living and host-associated isolates in terms of intergenic and intron streamlining.

We additionally find evidence consistent with host adaptation within *Blastocystis*. The *Blastocystis* strain isolated from an herbivorous tortoise contained many carbohydrate digesting enzymes not present in other *Blastocystis* isolates (Figure 3). In several cases, these genes were acquired via horizontal transfer from bacteria. The presence of plant-specific carbohydrate digesting enzymes suggests that, like some ciliates, *Blastocystis* may ferment plant carbohydates.

We find additional changes in *Blastocystis* genomes that likely reflect host interactions including the variable expansion of GT49 family glycosyltransferases (Figure 5). This family of glycosyltransferases has also expanded in the Mucormycota fungi, where they are hypothesized to participate in glycocalyx formation[70]. *Blastocystis* cells have a dense extracellular surface coat that is heavily glycosylated[71], and this coat has been observed to trap debris and other microbes which may play a role in obtaining nutrients[35,36]. Further, all *Blastocystis* isolates and *P. lacertae* contained large numbers of an apparently Opalinata-restricted gene family that resembles cadherins. Cadherins mediate cell-cell adhesion and interactions in animals; though similar proteins have been described in protists, their functions in protists are not well understood. The presence on the cell surface and large number of this unique gene family in both *Proteromonas*

and *Blastocystis* suggest a role in microbial or host interaction within the gut. These proteins may potentially provide anchors for adhering to host mucin or to epithelial cells. Alternatively, as these proteins are highly genetically diverse and taxonomically restricted, these proteins may function as antigens. Finally, *Blastocystis* isolates and *P. lacertae* all contained high copy numbers of C1A and C13 secreted cysteine proteases. Secreted cysteine proteases play diverse and important roles in host-parasite interactions. C1A cysteine proteases are common in pathogenic intestinal protists such as Entamoeba, where they play many roles including cleaving host mucus and enabling cell invasion[75,76]. C13 proteases are not frequently present in high copy numbers in most other gut eukaryotic protists, and their function in these organisms is not well understood. These cysteine protease genes were previously reported to be present in multiple human *Blastocystis* subtypes[31], but analyzing the additional tortoise *Blastocystis* genome, the *P. lacertae* genome, and the *C. burkhardae* genome demonstrates that these cysteine proteases are shared in similar copy numbers between *Blastocystis* and *P. lacertae*, but are absent from *C. burkhardae*. As these genes are maintained in similar copy numbers across these gut-dwelling organisms, it is likely that they play important roles in the gut environment.

Finally, all *Blastocystis* isolates sequenced here have lost genes associated with the more complex flagellated cell morphology common to other stramenopiles. *Blastocystis* isolates have lost nearly all cilia and flagellal related proteins, ion channel proteins, as well as substantially reduced the number of microtubule motor proteins (Figure 6). This universal loss suggests that the loss of cilia and flagellal proteins, along with the microtubule motors required both for movement and intraflagellar transport, is ancestral to the *Blastocystis* genus.

While *Blastocystis* is widespread across vertebrates, there is limited evidence for co-phylogeny of *Blastocystis* subtypes with their host (Figure 1). Instead, the *Blastocystis* phylogeny more closely resembles that of spore-forming gut bacteria[77], with host switching as a common phenomenon. Because of the lack of strong co-phylogenetic signal, a number of open questions remain about *Blastocystis* host association, including what the barriers to colonizing different hosts are and how multiple strains or subtypes (commonly termed 'mixed infections') are maintained, will require a more dense sampling of genomes across the *Blastocystis* tree of life. As human associated *Blastocystis* subtypes are not monophyletic, understanding the biology of the protists that impact humans will require considering *Blastocystis* subtypes that are present in the guts of diverse vertebrate hosts.

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, and a Benioff Center for Microbiome Medicine Trainee Pilot Award 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.

Data and code availability

Genome assemblies and annotations are pending publication in Genbank, and are currently available on Figshare at https://figshare.com/projects/Blastocystis_genomes/186136. RNA sequencing for all strains is available from the SRA under BioProject PRJNA1041441. Code to calculate the alien index is available at https://github.com/allind/alienindex. All other code used in this study is available at https://github.com/allind/blastannotate.

Figure legends

Figure 1. (A) 18S rRNA phylogeny of *Blastocystis* subtypes, with *Proteromonas lacertae* as an outgroup. (B) Schematic of relationships between organisms analyzed in this work. Important evolutionary transitions are highlighted.

Figure 2. (A) Element enriched downstream of genes in ST1, ST3, and ST4. Motif generated with STREME[78]. (B-D) Comparisons of genomic features across all *Blastocystis* strains, *P. lacertae*, and *C. burkhardae:* (B) Percentage of predicted genes with downstream conserved element, and percentage supported by RNA-seq evidence without stop codons in all *Blastocystis* strains. (C) Log-transformed intron lengths. (D) Intergenic lengths.

Figure 3. Gene families that are variable across *Blastocystis* isolates.

Figure 4. (A) Estimated time of transfer for previously identified horizontally transferred genes in *Blastocystis* ST1. (B) Gene tree of *Blastocystis* tortoise isolate GH154 gene and 100 bacterial homologues. Sequences highlighted in red are Bacteroidota sequences. (C) Gene tree of *Blastocystis* tortoise isolate alpha-L-rhamnosidase and 100 bacterial homologues. Sequences highlighted in blue are Bacillota sequences.

Figure 5. Gene families present in *P. lacerate* and all *Blastocystis* isolates but absent from *C. burkhardae*.

Figure 6. Loss of morphological genes in *Blastocystis.*

Figure S1. Synteny of *Blastocystis* genomes based on reciprocal best blast hits. Plot created with ODP[79].

Figure S2. Average percent identity across 1,350 single-copy orthologs for *Blastocystis*.

Figure S3. (A) Intergenic region of two genes encoded on opposite strands, one containing a canonical DNA stop codon and one lacking a stop codon. Tracks represent aligned RNA-seq coverage and exon-exon junctions. (B) The conserved polyadenylation motif is immediately downstream of the gene lacking the stop codon.

Figure S4. (A) Number of introns per 1,350 single-copy ortholog in all *Blastocystis* isolates. Vertical line depicts median number of introns per gene. (B) Log-transformed length distribution

of introns in 1,350 single-copy orthologs in all *Blastocystis* isolates. Vertical line depicts median intron size.

Figure S5. Gene trees of horizontally transferred genes in the tortoise *Blastocystis* isolate. Gene trees were constructed with IQ-TREE and midpoint rooted. Branches with bootstrap support less than 70 are collapsed. (A) Formate-nitrate transporter gene tree. (B) DUF711-containing gene tree. (C) GH154 gene tree. (D) Beta-D-xylosidase gene tree. (E) Alpha-L-rhamnosidase gene tree. (F) GH88 gene tree.

Table S1. Xenic bacterial contigs identified with GTDB.

Table S2. GenBank accessions of 18S rRNA sequences used to construct *Blastocystis* phylogeny.

Table S3. Genes containing InterPro domains per genome assembly, annotated by variability.

References

1. Yadav, M., Verma, M. K. & Chauhan, N. S. A review of metabolic potential of human gut microbiome in human nutrition. *Arch. Microbiol.* **200**, 203–217 (2018).
2. Sommer, F. & Bäckhed, F. The gut microbiota — masters of host development and physiology. *Nat. Rev. Microbiol.* **11**, 227–238 (2013).
3. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
4. Thorman, A. W. *et al.* Gut Microbiome Composition and Metabolic Capacity Differ by FUT2 Secretor Status in Exclusively Breastfed Infants. *Nutrients* **15**, 471 (2023).
5. Godoy-Vitorino, F. *et al.* Comparative analyses of foregut and hindgut bacterial communities in hoatzins and cows. *ISME J.* **6**, 531–541 (2012).
6. Campo, J. del, Bass, D. & Keeling, P. J. The eukaryome: Diversity and role of microeukaryotic organisms associated with animal hosts. *Funct. Ecol.* **34**, 2045–2054 (2020).
7. Howitt, M. R. *et al.* Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science* **351**, 1329–1333 (2016).

8. Brugerolle, G. & Radek, R. Symbiotic Protozoa of Termites. in *Intestinal Microorganisms of Termites and Other Invertebrates* (eds. König, H. & Varma, A.) 243–269 (Springer, 2006). doi:10.1007/3-540-28185-1_10.

9. Li, Z. *et al.* Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates. *ISME J.* **16**, 2775–2787 (2022).

10. Scanlan, P. D. & Stensvold, C. R. Blastocystis: getting to grips with our guileful guest. *Trends Parasitol.* **29**, 523–529 (2013).

11. Scanlan, P. D. & Marchesi, J. R. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME J.* **2**, 1183–1193 (2008).

12. Scanlan, P. D. *et al.* The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol. Ecol.* **90**, 326–330 (2014).

13. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).

14. Gantois, N. *et al.* First Report on the Prevalence and Subtype Distribution of Blastocystis sp. in Edible Marine Fish and Marine Mammals: A Large Scale-Study Conducted in Atlantic Northeast and on the Coasts of Northern France. *Microorganisms* **8**, 460 (2020).

15. Boreham, P. F. L. & Stenzel, D. J. Blastocystis in Humans and Animals: Morphology, Biology, and Epizootiology. in *Advances in Parasitology* (eds. Baker, J. R. & Muller, R.) vol. 32 1–70 (Academic Press, 1993).

16. Yoshikawa, H., Koyama, Y., Tsuchiya, E. & Takami, K. Blastocystis phylogeny among various isolates from humans to insects. *Parasitol. Int.* **65**, 750–759 (2016).

17. Stensvold, C. R. *et al.* Terminology for Blastocystis subtypes--a consensus. *Trends Parasitol.* **23**, 93–96 (2007).

18. Kostka, M. Opalinata. in *Handbook of the Protists* (eds. Archibald, J. M., Simpson, A. G. B. & Slamovits, C. H.) 543–565 (Springer International Publishing, 2017). doi:10.1007/978-3-319-28149-0_4.

19. Zierdt, C. H. Blastocystis hominis, a long-misunderstood intestinal parasite. *Parasitol. Today Pers. Ed* **4**, 15–17 (1988).

20. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

21. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

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

23. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLOS Comput. Biol.* **18**, e1009802 (2022).

24. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).

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

26. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

27. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569–580 (2018).
28. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
29. Gerrick, E. R. *et al.* Commensal protists in reptiles display flexible host range and adaptation to ectothermic hosts. *mBio* **0**, e02273-23 (2023).
30. Klimeš, V., Gentekaki, E., Roger, A. J. & Eliáš, M. A Large Number of Nuclear Genes in the Human Parasite Blastocystis Require mRNA Polyadenylation to Create Functional Termination Codons. *Genome Biol. Evol.* **6**, 1956–1961 (2014).
31. Gentekaki, E. *et al.* Extreme genome diversity in the hyper-prevalent parasitic eukaryote Blastocystis. *PLOS Biol.* **15**, e2003769 (2017).
32. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, giy093 (2018).
33. Gabriel, L. *et al.* BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *BioRxiv Prepr. Serv. Biol.* 2023.06.10.544449 (2023) doi:10.1101/2023.06.10.544449.
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
35. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

36. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2022).

37. Záhonová, K. *et al.* Evolutionary analysis of cellular reduction and anaerobicity in the hyper-prevalent gut microbe Blastocystis. *Curr. Biol. CB* **33**, 2449-2464.e8 (2023).

38. Hackl, T. *et al.* Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate Cafeteria roenbergensis. *Sci. Data* **7**, 29 (2020).

39. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

40. Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225-229 (2011).

41. Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H. & Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* **50**, W228–W234 (2022).

42. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

43. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

44. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

45. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

46. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210–1213 (2008).

47. Wisecaver, J. H., Alexander, W. G., King, S. B., Todd Hittinger, C. & Rokas, A. Dynamic Evolution of Nitric Oxide Detoxifying Flavohemoglobins, a Family of Single-Protein Metabolic Modules in Bacteria and Eukaryotes. *Mol. Biol. Evol.* **33**, 1979–1987 (2016).

48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

49. Stensvold, C. R. & Clark, C. G. Pre-empting Pandora's Box: Blastocystis Subtypes Revisited. *Trends Parasitol.* **36**, 229–232 (2020).

50. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

51. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

52. Hu, E.-Z., Lan, X.-R., Liu, Z.-L., Gao, J. & Niu, D.-K. A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics* **23**, 110 (2022).

53. Aliperti, L. *et al.* r/K selection of GC content in prokaryotes. *Environ. Microbiol.* **n/a**,.

54. Denoeud, F. *et al.* Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite. *Genome Biol.* **12**, R29 (2011).

55. De Lorenzo, G. & Ferrari, S. Polygalacturonase-inhibiting proteins in defense against phytopathogenic fungi. *Curr. Opin. Plant Biol.* **5**, 295–299 (2002).

56. Deng, Q. *et al.* Characterization of Two Novel Rumen-Derived Exo-Polygalacturonases: Catalysis and Molecular Simulations. *Microorganisms* **11**, 760 (2023).

57. Yang, J. K. *et al.* Crystal structure of beta-D-xylosidase from Thermoanaerobacterium saccharolyticum, a family 39 glycoside hydrolase. *J. Mol. Biol.* **335**, 155–165 (2004).

58. Stevens, C. E. & Hume, I. D. *Comparative physiology of the vertebrate digestive system*. (Cambridge University Press, 1995).

59. Gaythri, T., Suresh, K., Subha, B. & Kalyani, R. Identification and Characterisation of Heat Shock Protein 70 in Thermal Stressed Blastocystis sp. *PLoS ONE* **9**, e95608 (2014).

60. Day, J., Passecker, A., Beck, H.-P. & Vakonakis, I. The Plasmodium falciparum Hsp70-x chaperone assists the heat stress response of the malaria parasite. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **33**, 14611–14624 (2019).

61. Eme, L., Gentekaki, E., Curtis, B., Archibald, J. M. & Roger, A. J. Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut. *Curr. Biol. CB* **27**, 807–820 (2017).

62. Cartmell, A. *et al.* A surface endogalactanase in Bacteroides thetaiotaomicron confers keystone status for arabinogalactan degradation. *Nat. Microbiol.* **3**, 1314–1326 (2018).

63. Stairs, C. W. *et al.* Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr. Biol. CB* **31**, 5605-5612.e5 (2021).

64. Wells, L. The O-Mannosylation Pathway: Glycosyltransferases and Proteins Implicated in Congenital Muscular Dystrophy. *J. Biol. Chem.* **288**, 6930–6935 (2013).

65. Khan, S. & Scholey, J. M. Assembly, Functions and Evolution of Archaella, Flagella and Cilia. *Curr. Biol. CB* **28**, R278–R292 (2018).

66. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2012).

67. Leckenby, A. & Hall, N. Genomic changes during evolution of animal parasitism in eukaryotes. *Curr. Opin. Genet. Dev.* **35**, 86–92 (2015).

68. Wang, S. E., Amir, A. S., Nguyen, T., Poole, A. M. & Simoes-Barbosa, A. Spliceosomal introns in Trichomonas vaginalis revisited. *Parasit. Vectors* **11**, 607 (2018).

69. Wilihoeft, U., Campos-Góngora, E., Touzni, S., Bruchhaus, I. & Tannich, E. Introns of Entamoeba histolytica and Entamoeba dispar. *Protist* **152**, 149–156 (2001).

70. Orłowska, M. & Muszewska, A. In Silico Predictions of Ecological Plasticity Mediated by Protein Family Expansions in Early-Diverging Fungi. *J. Fungi* **8**, 67 (2022).

71. Lanuza, M. D., Carbajal, J. A. & Borrás, R. Identification of surface coat carbohydrates in Blastocystis hominis by lectin probes. *Int. J. Parasitol.* **26**, 527–532 (1996).

72. Tan, K. S. W. New Insights on Classification, Identification, and Clinical Relevance of Blastocystis spp. *Clin. Microbiol. Rev.* **21**, 639–665 (2008).

73. Yason, J. A. & Tan, K. S. W. Membrane Surface Features of Blastocystis Subtypes. *Genes* **9**, 417 (2018).

74. Zaman, V., Howe, J. & Ng, M. Observations on the surface coat of Blastocystis hominis. *Parasitol. Res.* **83**, 731–733 (1997).

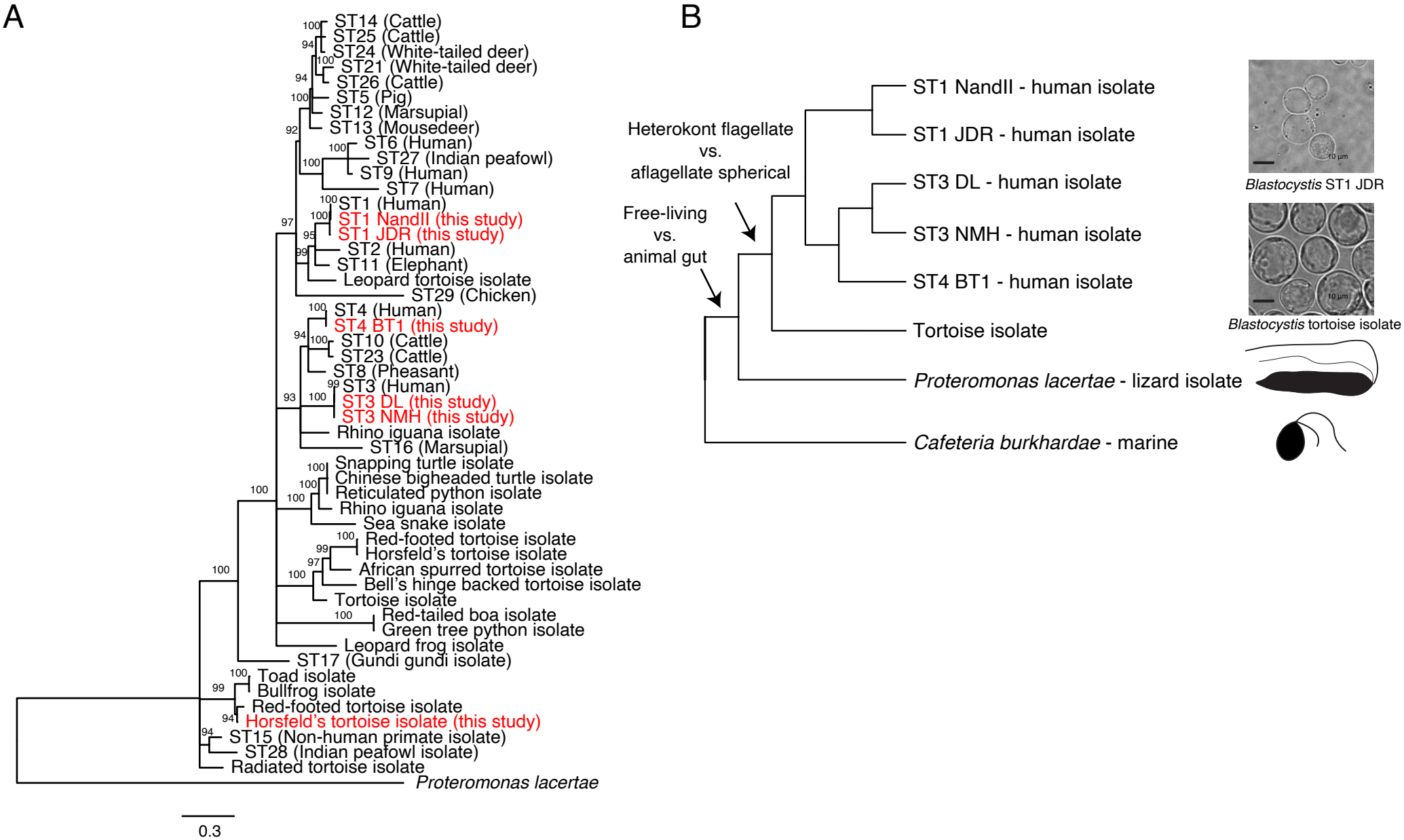
75. Lidell, M. E., Moncada, D. M., Chadee, K. & Hansson, G. C. Entamoeba histolytica cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel. *Proc. Natl. Acad. Sci.* **103**, 9298–9303 (2006).

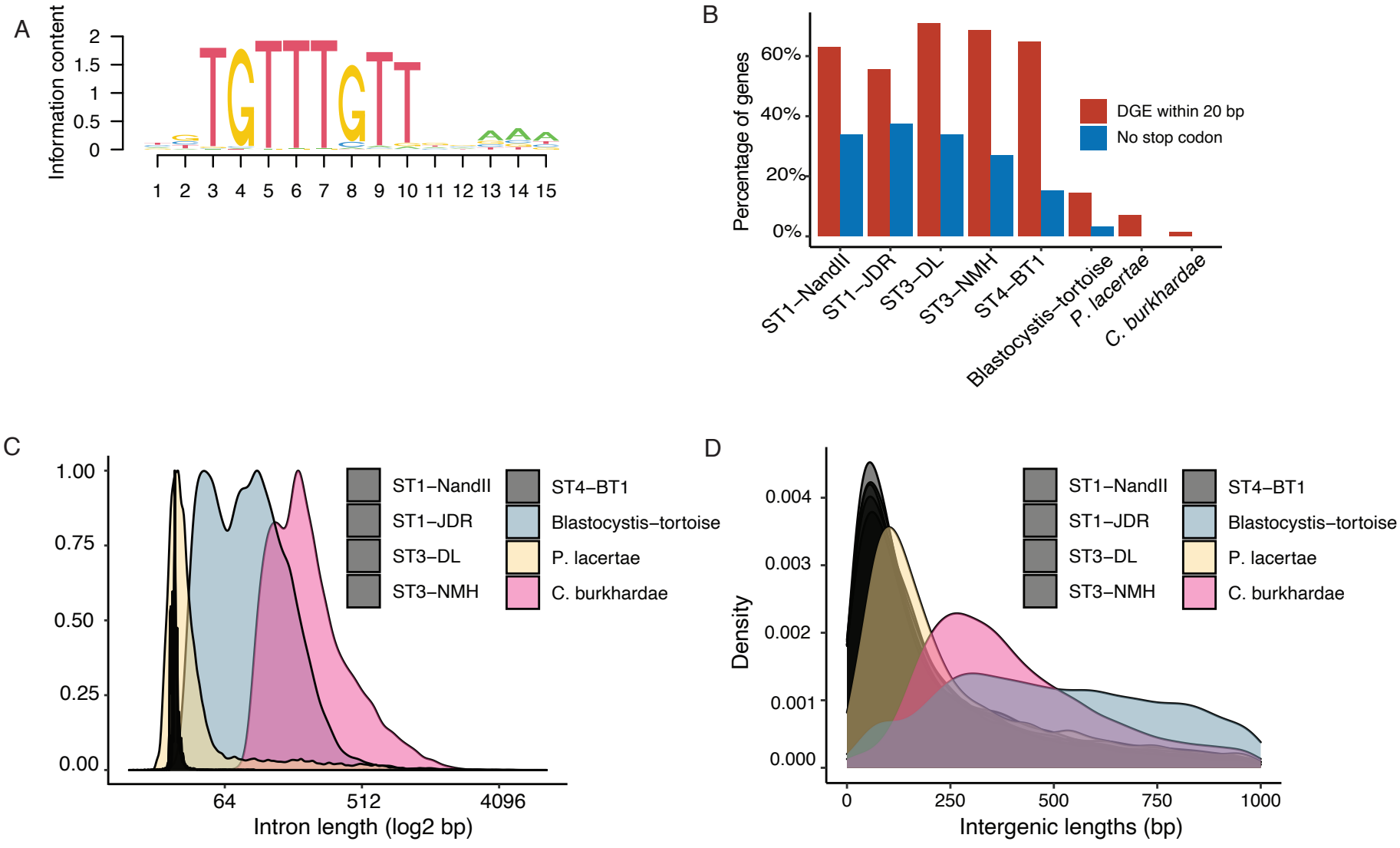
76. Que, X. & Reed, S. L. Cysteine Proteinases and the Pathogenesis of Amebiasis. *Clin. Microbiol. Rev.* **13**, 196–206 (2000).

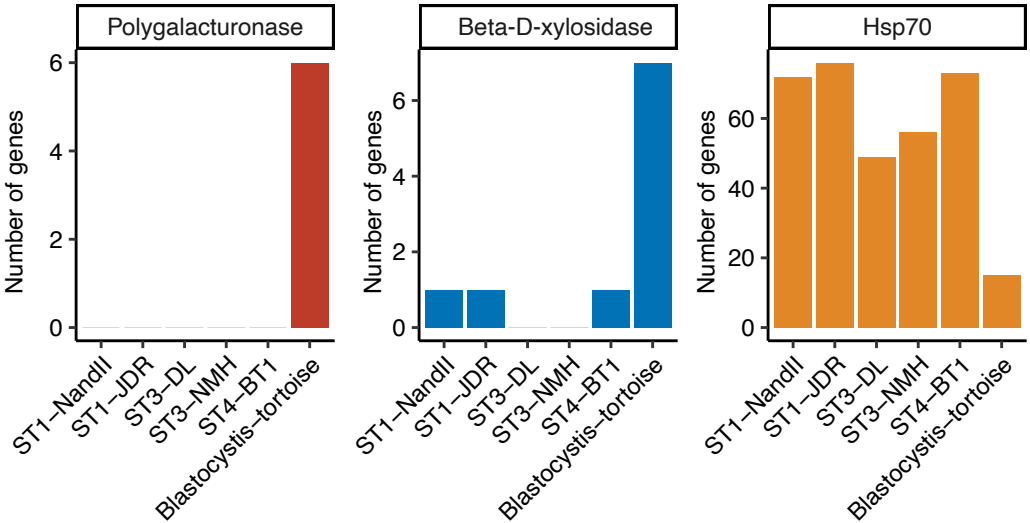
77. Moeller, A. H. *et al.* Cospeciation of gut microbiota with hominids. *Science* **353**, 380–382 (2016).

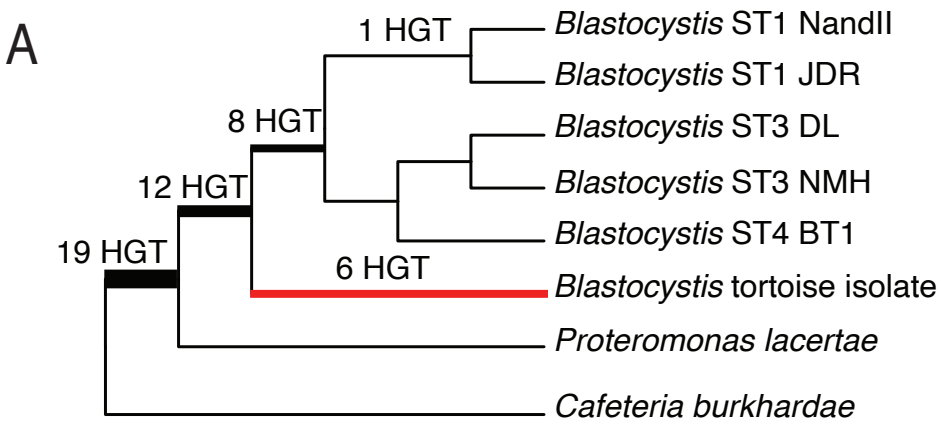
78. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinforma. Oxf. Engl.* **37**, 2834–2840 (2021).

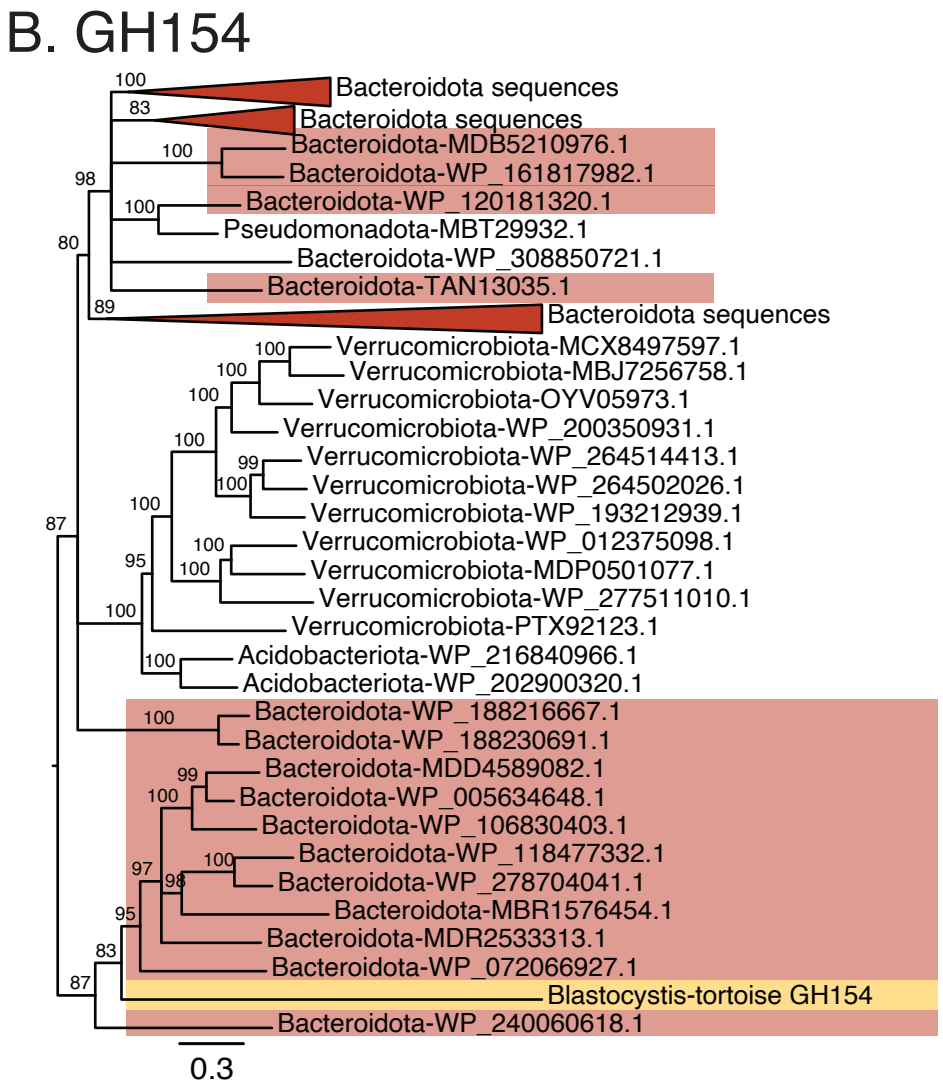
79. Schultz, D. T. *et al.* Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).

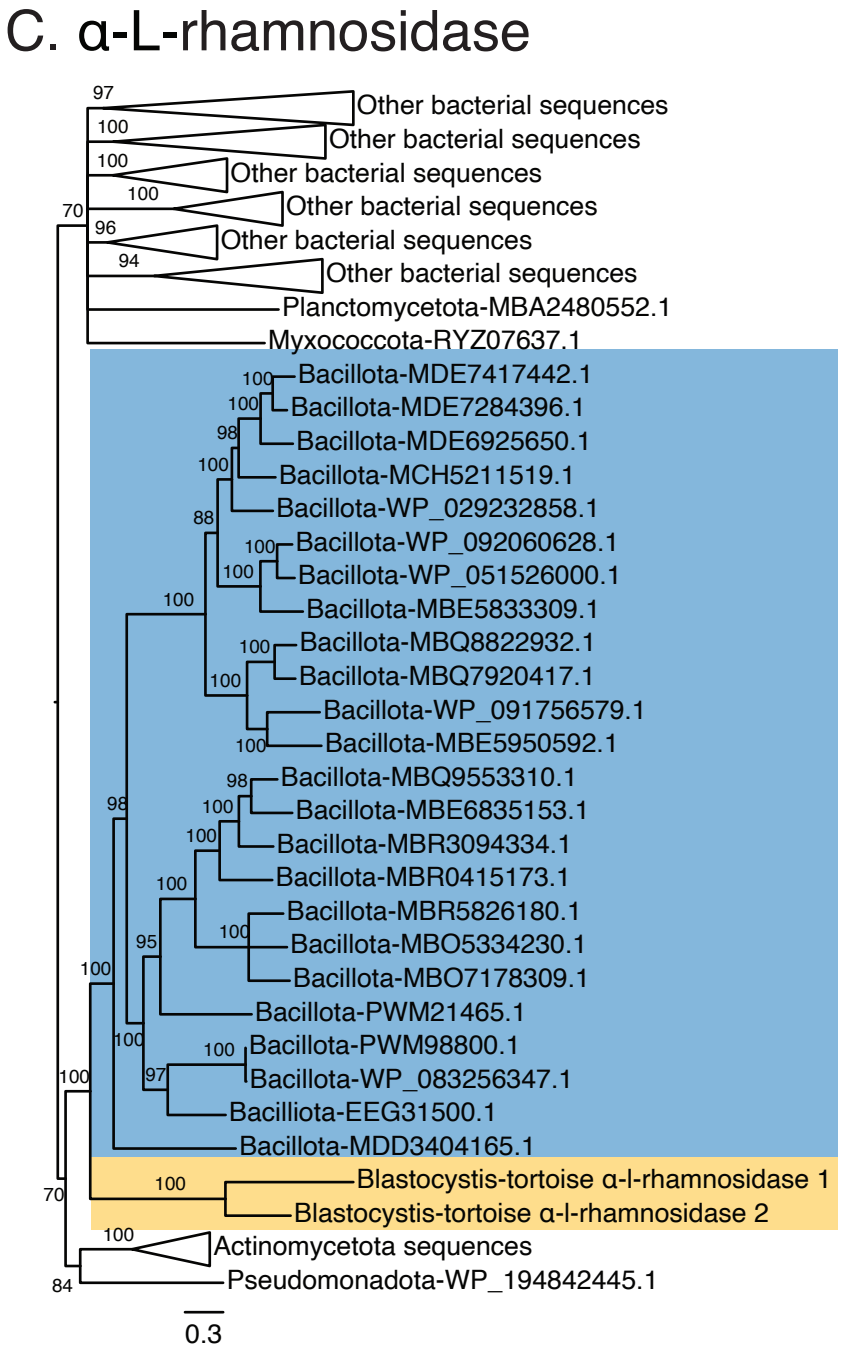

A
ST14 (Cattle)
ST25 (Cattle)
ST24 (White-tailed deer)
ST21 (White-tailed deer)
ST26 (Cattle)
ST5 (Pig)
ST12 (Marsupial)
ST13 (Mousedeer)
ST6 (Human)
ST27 (Indian peafowl)
ST9 (Human)
ST7 (Human)
ST1 (Human)
ST1 NandII (this study)
ST1 JDR (this study)
ST2 (Human)
ST11 (Elephant)
Leopard tortoise isolate
ST29 (Chicken)
ST4 (Human)
ST4 BT1 (this study)
ST10 (Cattle)
ST23 (Cattle)
ST8 (Pheasant)
ST3 (Human)
ST3 DL (this study)
ST3 NMH (this study)
Rhino iguana isolate
ST16 (Marsupial)
Snapping turtle isolate
Chinese bigheaded turtle isolate
Reticulated python isolate
Rhino iguana isolate
Sea snake isolate
Red-footed tortoise isolate
Horsfeld's tortoise isolate
African spurred tortoise isolate
Bell's hinge backed tortoise isolate
Tortoise isolate
Red-tailed boa isolate
Green tree python isolate
Leopard frog isolate
ST17 (Gundi gundi isolate)
Toad isolate
Bullfrog isolate
Red-footed tortoise isolate
Horsfeld's tortoise isolate (this study)
ST15 (Non-human primate isolate)
ST28 (Indian peafowl isolate)
Radiated tortoise isolate
Proteromonas lacertae
0.3
B
Heterokont flagellate
vs.
aflagellate spherical
Free-living
vs.
animal gut
ST1 NandII - human isolate
ST1 JDR - human isolate
ST3 DL - human isolate
ST3 NMH - human isolate
ST4 BT1 - human isolate
Tortoise isolate
Proteromonas lacertae - lizard isolate
Cafeteria burkhardae - marine
Blastocystis ST1 JDR
Blastocystis tortoise isolate

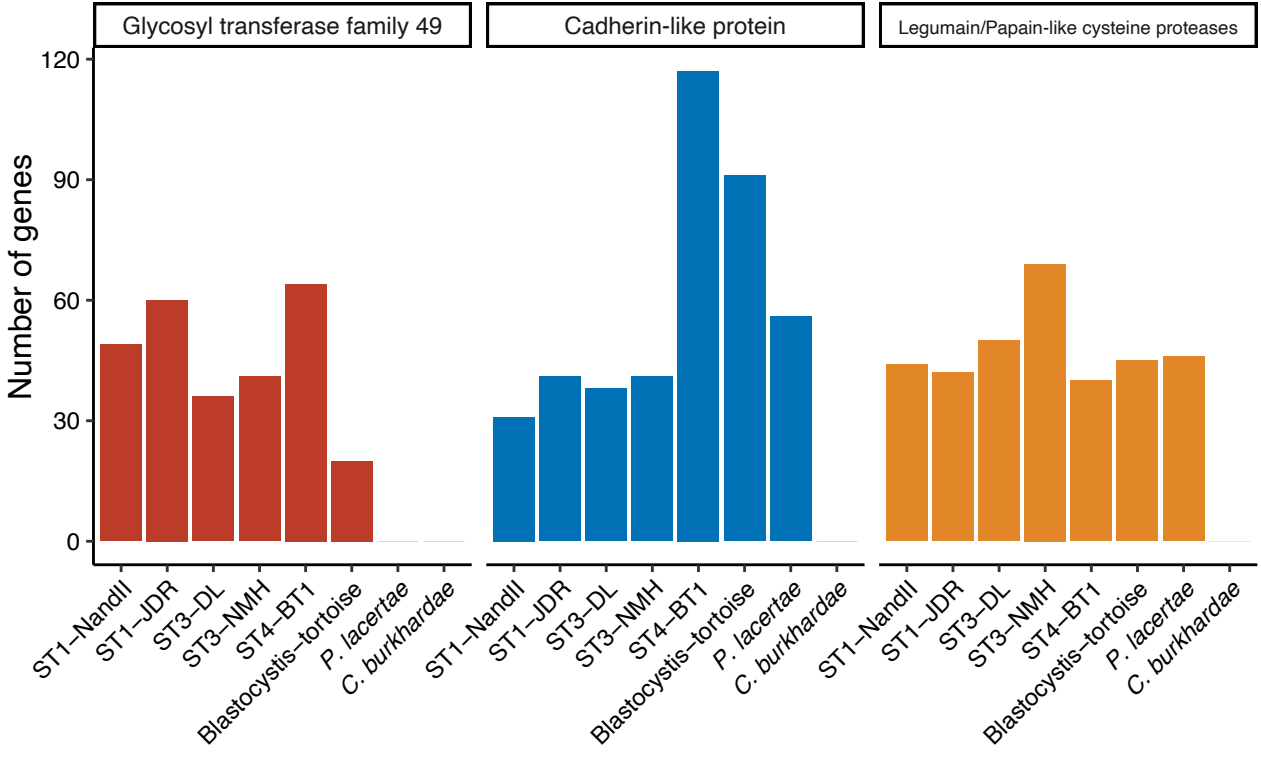

A
Information content
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
B
Percentage of genes
60%
40%
20%
0%
DGE within 20 bp
No stop codon
ST1–NandII
ST1–JDR
ST3–DL
ST3–NMH
ST4–BT1
Blastocystis–tortoise
P. lacertae
C. burkhardae
C
1.00
0.75
0.50
0.25
0.00
64
512
4096
Intron length (log2 bp)
ST1–NandII
ST1–JDR
ST3–DL
ST3–NMH
ST4–BT1
Blastocystis–tortoise
P. lacertae
C. burkhardae
D
Density
0.004
0.003
0.002
0.001
0.000
0
250
500
750
1000
Intergenic lengths (bp)
ST1–NandII
ST1–JDR
ST3–DL
ST3–NMH
ST4–BT1
Blastocystis–tortoise
P. lacertae
C. burkhardae

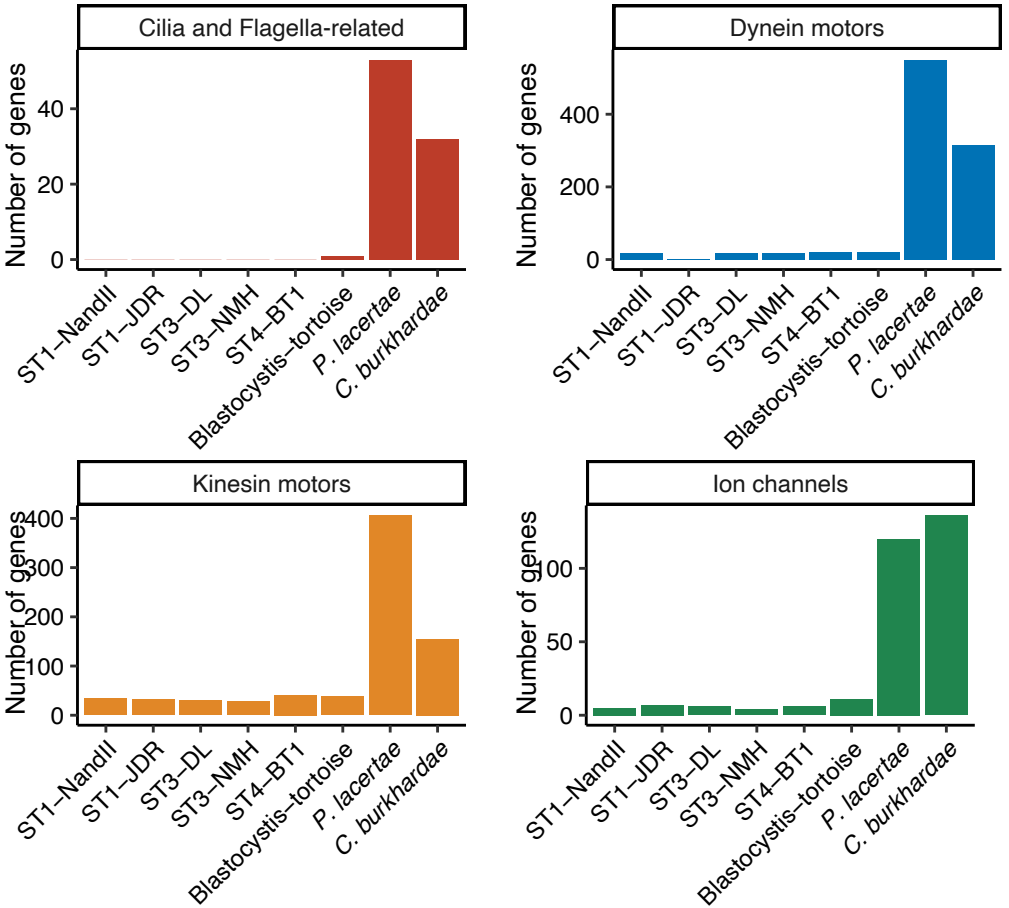

Polygalacturonase
Beta-D-xylosidase
Hsp70
Number of genes
0
2
4
6
20
40
60
ST1–NandII
ST1–JDR
ST3–DL
ST3–NMH
ST4–BT1
Blastocystis–tortoise


A
1 HGT
8 HGT
12 HGT
19 HGT
6 HGT
Blastocystis ST1 NandII
Blastocystis ST1 JDR
Blastocystis ST3 DL
Blastocystis ST3 NMH
Blastocystis ST4 BT1
Blastocystis tortoise isolate
Proteromonas lacertae
Cafeteria burkhardae


B. GH154
Bacteroidota sequences
Bacteroidota sequences
Bacteroidota-MDB5210976.1
Bacteroidota-WP_161817982.1
Bacteroidota-WP_120181320.1
Pseudomonadota-MBT29932.1
Bacteroidota-WP_308850721.1
Bacteroidota-TAN13035.1
Bacteroidota sequences
Verrucomicrobiota-MCX8497597.1
Verrucomicrobiota-MBJ7256758.1
Verrucomicrobiota-OYV05973.1
Verrucomicrobiota-WP_200350931.1
Verrucomicrobiota-WP_264514413.1
Verrucomicrobiota-WP_264502026.1
Verrucomicrobiota-WP_193212939.1
Verrucomicrobiota-WP_012375098.1
Verrucomicrobiota-MDP0501077.1
Verrucomicrobiota-WP_277511010.1
Verrucomicrobiota-PTX92123.1
Acidobacteriota-WP_216840966.1
Acidobacteriota-WP_202900320.1
Bacteroidota-WP_188216667.1
Bacteroidota-WP_188230691.1
Bacteroidota-MDD4589082.1
Bacteroidota-WP_005634648.1
Bacteroidota-WP_106830403.1
Bacteroidota-WP_118477332.1
Bacteroidota-WP_278704041.1
Bacteroidota-MBR1576454.1
Bacteroidota-MDR2533313.1
Bacteroidota-WP_072066927.1
Blastocystis-tortoise GH154
Bacteroidota-WP_240060618.1
0.3


C. α-L-rhamnosidase
Other bacterial sequences
Other bacterial sequences
Other bacterial sequences
Other bacterial sequences
Other bacterial sequences
Other bacterial sequences
Planctomycetota-MBA2480552.1
Myxococcota-RYZ07637.1
Bacillota-MDE7417442.1
Bacillota-MDE7284396.1
Bacillota-MDE6925650.1
Bacillota-MCH5211519.1
Bacillota-WP_029232858.1
Bacillota-WP_092060628.1
Bacillota-WP_051526000.1
Bacillota-MBE5833309.1
Bacillota-MBQ8822932.1
Bacillota-MBQ7920417.1
Bacillota-WP_091756579.1
Bacillota-MBE5950592.1
Bacillota-MBQ9553310.1
Bacillota-MBE6835153.1
Bacillota-MBR3094334.1
Bacillota-MBR0415173.1
Bacillota-MBR5826180.1
Bacillota-MBO5334230.1
Bacillota-MBO7178309.1
Bacillota-PWM21465.1
Bacillota-PWM98800.1
Bacillota-WP_083256347.1
Bacilliota-EEG31500.1
Bacillota-MDD3404165.1
Blastocystis-tortoise α-l-rhamnosidase 1
Blastocystis-tortoise α-l-rhamnosidase 2
Actinomycetota sequences
Pseudomonadota-WP_194842445.1
0.3


Glycosyl transferase family 49
Cadherin-like protein
Legumain/Papain-like cysteine proteases
Number of genes
120
90
60
30
0
ST1–NandII
ST1–JDR
ST3–DL
ST3–NMH
ST4–BT1
Blastocystis–tortoise
P. lacertae
C. burkhardae


Cilia and Flagella-related
Dynein motors
Kinesin motors
Ion channels
Number of genes
0
20
40
200
400
100
50
ST1–NandII
ST1–JDR
ST3–DL
ST3–NMH
ST4–BT1
Blastocystis–tortoise
P. lacertae
C. burkhardae