Stensvold Christen Rune (Orcid ID: 0000-0002-1417-7048)
El-Badry Ayman A. (Orcid ID: 0000-0002-9673-2622)
van der Giezen Mark (Orcid ID: 0000-0002-1033-1335)
Clark C Graham (Orcid ID: 0000-0002-0521-0977)

Stensvold et al. – On *E. coli* and *E. hartmanni* diversity

# Further Insight into the Genetic Diversity of *Entamoeba coli* and *Entamoeba hartmanni*

Christen Rune Stensvold[a], Kasandra Ascuña-Durand[b], Amal Chihi[c], Salem Belkessa[d], Özgür Kurt[e], Ayman El-Badry[f], Mark van der Giezen[g], Graham Clark[h]

**Authors' addresses:**

a) Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen S, Denmark.
b) Laboratorio de Microbiología Molecular, Facultad de Medicina, Universidad Nacional de San Agustín, 04001 Arequipa, Peru.
c) Laboratoire de Recherche 'Parasitologie Médicale, Biotechnologies et Biomolécules', LR 16-IPT-06, Université Tunis El-Manar, Institut Pasteur de Tunis, 13 place Pasteur, B.P. 74 1002, Tunis Belvédère, Tunisia.
d) Department of Biology, Faculty of Nature and Life Sciences, Ziane Achour University of Djelfa, Moudjbara Road, BP 3117, Djelfa 17000, Algeria.
e) Department of Medical Microbiology, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Atasehir, Istanbul 34770, Turkey.
f) Department of Microbiology, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam 31451, Saudi Arabia.
g) Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway.
h) Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

**Corresponding author:** C. R. Stensvold, Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen S, Denmark.
e-mail: run@ssi.dk

## ABSTRACT

Despite the species' wide distribution, studies of the genetic diversity within *Entamoeba coli* and *Entamoeba hartmanni* remain limited. In the present study, we provide further insight into the genetic diversity of both species based on analysis of partial nuclear small subunit ribosomal DNA sequences generated from human faecal DNAs from samples collected in Africa, South America and Europe. Reinforcing the previous recognition that *E. coli* is a species complex, our data confirm the existence of the two subtypes, ST1 and ST2, previously identified plus, potentially, a new subtype, ST3. While ST1 appears to be genetically quite homogenous, ST2 shows a substantial degree of intra-subtype diversity. ST2 was more common in samples collected outside Europe, whereas ST1 showed no geographical restriction. The potentially novel subtype is represented to date exclusively by sequences from South American and African samples. In contrast to previous reports, our new data also indicate substantial variation in *E. hartmanni* that could also support the establishment of subtypes within this species. Here, however, no links were identified between subtype and geographical origin.

**Keywords:** co-evolution; DNA-based taxonomy; evolution; host specificity; protists; endobionts; ribosomal DNA; taxonomy.

*ENTAMOEBA coli* and *Entamoeba hartmanni* are both considered nonpathogenic endobionts of the gastrointestinal tract. The life cycles of both species involve trophozoite and cyst stages, both of which may be observed in faecal samples. Despite their high prevalence and cosmopolitan distribution, studies of the genetic diversity and host specificity of these two species remain limited.

Mature *E. coli* cysts, which can attain sizes of up to 30–35 µm, typically contain eight nuclei, a classic morphological hallmark that has been used to identify *E. coli* in faecal concentrates and differentiate it from other species of *Entamoeba* found in humans. *Entamoeba coli* appears to be an endobiont mainly of primates, but possibly also colonises other mammals. Based on morphology, *E. coli* has also been reported from, among others, pigs (Pinilla et al. 2021) and dogs (Campos Filho et al. 2008). In addition to *E. coli*, octonucleate cyst-producing species of *Entamoeba* for which DNA sequences are available include *Entamoeba muris* (Kobayashi et al. 2009) and *Entamoeba* sp. RL7 (Jacob et al. 2016).

*Entamoeba hartmanni* is also an endobiont of both humans and non-human primates, and this species has been identified recently in pigs via metabarcoding of faecal samples (Stensvold et al. 2021b). Mature cysts contain four nuclei, but, with a maximum size of 10 µm, cysts of *E. hartmanni* are typically distinctly smaller than those of other species that produce quadrinucleated cysts, such as *Entamoeba histolytica* and *Entamoeba dispar*.

In 2011, extensive cryptic sequence diversity in the small subunit ribosomal RNA (*SSU rRNA*) genes was identified within *E. coli*, leading to the recognition of two distinct subtypes (ST), ST1 and ST2, and indicating that *E. coli* is a species complex (Stensvold et al. 2011). This situation is similar to that found in *Entamoeba gingivalis* (Stensvold et al. 2021a, Cembranelli et al. 2013, Garcia et al. 2018), *Entamoeba moshkovskii* (Stensvold et al., 2020), and *Entamoeba polecki* (Verweij et al. 2001, Stensvold et al. 2018, Stensvold et al. 2011), as well as other archaemoebid species such as *Iodamoeba bütschlii* and *Endolimax nana* (Stensvold et al. 2012, Stensvold et al. 2020, Poulsen and Stensvold 2016). Since their description, these two *E. coli* STs have been confirmed in independent studies (Chihi et al. 2022, Chihi et al. 2019, Calegar et al. 2021, Stensvold et al. 2020, Jirků-Pomajbíková et al. 2016). Nevertheless, further sampling is necessary to evaluate whether the diversity discovered so far provides the full picture of the variation existing within this species.

For *E. hartmanni*, in contrast, SSU rDNA genetic diversity sufficient to justify splitting the species into subtypes was not found (Stensvold et al. 2011), and additional data produced recently (Calegar et al. 2021) did not challenge this conclusion.

Metabarcoding represents a method that can be used to detect and differentiate SSU rDNA sequences in complex samples. Although the assay is generally less sensitive than other DNA-based methods, such as specific real-time PCR assays, it has proven a useful tool for screening for selected parasites in one of our laboratories (Stensvold et al. 2020, Stensvold et al. 2021b, Holmgaard et al. 2020, Krogsgaard et al. 2018, Hartmeyer et al. 2019). The method enables extraction of sequences for all non-viral organisms present in any given sample.

The aim of this study was to provide further insight into the genetic diversity of *E. coli* and *E. hartmanni* by screening SSU rDNA sequence data generated by metabarcoding of 1,457 human faecal samples collected on three continents.

## MATERIALS AND METHODS

### Data

The metabarcoding assay used has been described in detail previously (Stensvold et al. 2020, Stensvold et al. 2021b, Krogsgaard et al. 2018). Briefly, three primer pairs of low specificity (G3, G4, and G6 primer pairs) were used to amplify nuclear SSU rDNA, and the PCR products were sequenced using an ILLUMINA MiSeq sequencer (ILLUMINA Inc, San Diego, CA). The resulting sequence reads were annotated to the lowest possible taxonomic level using the in-house software BION. Parts of the metabarcoding data used in the present study had been generated for various independent research projects and routine diagnostic purposes and were recently used in a separate study (Stensvold et al. 2022). Data representing faecal samples collected in Peru and Tunisia (Chihi et al. 2022, Chihi et al. 2019, Ascuna-Durand et al. 2020) were added. Thus, the 1,457 faecal DNA datasets represented populations sampled in Algeria, Denmark, Egypt, Peru, Tunisia, Turkey, and United Kingdom, with and without gastrointestinal symptoms. With respect to the local General Data Protection Regulation rules, the only metadata available for the samples were geographical origins. For sequence annotation purposes, sequences from samples from Algeria, Tunisia, and Egypt were marked with an ‘A’ (to denote Africa); sequences from Peru with an ‘SA’ (to denote South America), and sequences from Denmark, UK, and Turkey with an ‘E’ (to denote Europe). The number of samples screened from each region was 51, 119, 293, and 994 from Peru, Turkey, Northern Africa, and Europe, respectively.

Metabarcoding data for 1,457 human faecal DNAs were screened manually to identify samples positive for *E. coli* and *E. hartmanni.* From each positive sample, FASTA files specific to *E. coli* and *E. hartmanni* were retrieved and aligned using the Clustal Omega online sequence alignment tool (available at https://www.ebi.ac.uk/Tools/msa/clustalo/). Consensus sequences were generated manually by visual inspection in Jalview (available at https://www.jalview.org/).

Of the three primer pairs targeting eukaryotic DNA, the G4 primer pair was most efficient (i.e., generated more reads) at amplifying *E. coli*-specific DNA, and so, for consistency, the *E. coli* part of the study used sequence data produced by this primer pair only. However, the *E. coli*-specific PCR product generated by the G4 primer pair covers about 690 bp of the *SSU rRNA* gene, a product size that cannot be completely sequenced using ILLUMINA MiSeq sequencing technology (which only allows complete sequencing of products up to sizes of 500–600 bp (https://www.illumina.com/systems/sequencing-platforms.html). This means that each of the sequences had a gap in the middle (i.e., the forward and reverse sequences did not overlap). For *E. hartmanni*, the G6 primer pair was most efficient at generating *E. hartmanni*-specific sequences; for these sequences, no gap was present in the middle section of the sequences.

The sequences generated in the present study were submitted to GenBank with the accession numbers ON989959-ON990035 for *E. coli* and ON974211-ON974233 for *E. hartmanni.*

### Phylogenetic analysis

The part of the gene amplified for *E. coli* covered positions 630–1288 (with a sequence gap corresponding to positions 857-1041) relative to the reference sequence AF149914. The part of the gene amplified for *E. hartmanni* covered positions 566–980 (with no sequence gap) relative to the reference sequence AF149907.

For sequence alignment, consensus sequences from each sample were aligned with reference sequences from the NCBI Database spanning the sequenced regions (i.e., only reference sequences with 100% coverage of the amplified region were included). For the *E. coli* reference sequences, the region spanning the gap in the middle of the sequences was deleted from the alignment.

For both species, the manually edited sequence alignments were subject to phylogenetic analysis using the Neighbor-Joining algorithm as implemented in the MEGA7 software package (Kumar et al. 2016) with 1,000 bootstrap replicates.

To further inform the inferences from the data generated in the present study, *E. coli*- and *E. hartmanni*-specific sequences derived from the 5'-end of the *SSU rRNA* genes were also downloaded from GenBank, aligned and subjected to separate phylogenetic analysis.

## RESULTS

### *Entamoeba coli*

The entire dataset for *E. coli* included 88 partial SSU rDNA sequences, including 77 consensus sequences from this study and 11 reference sequences. The reference sequences were annotated as *E. coli* (eight sequences of which five were from humans, two from non-human primates, and one from a rat), *E. muris* from a Mongolian gerbil (AB445018), *Entamoeba* sp. RL7 from a Phayre's leaf monkey (FR686360), and *Entamoeba* sp. RL11 from a short-tailed field vole (KR025409). The last three sequences were included as an outgroup. The *E. coli* reference sequences were chosen to ensure representation of the known range of genetic diversity within the species while trying to avoid redundancy.

Of the 77 consensus sequences, 45 belonged to *E. coli* ST1, and 27 to *E. coli* ST2 (Fig 1). The remaining five sequences clustered with reference sequence *E. coli* S2702 (FR686364) in a distinct group with high bootstrap support (99%), which we propose to call 'ST3'; this appears to share a most recent common ancestor with ST2, although with only moderate bootstrap support. ST3 (sequence FR686364) differs from ST1 (AF149915) and ST2 (AF149914) by 7% and 11%, respectively.

Despite the fact that the SSU rDNA region studied was relatively small, and smaller than the region analysed in our previous study (Stensvold et al. 2011), the topology of the tree exhibited the same overall structure as identified previously; in short, ST1 sequences were very similar, whereas several subclusters were observed within ST2 (Fig 1). The presence of subclusters within ST2 was further supported by our supplementary analysis of the 5'-end region of *E. coli* SSU rDNA (Suppl Fig 1). In that analysis, *E. coli* sequences from gorillas obtained by Nolan and colleagues (Nolan et al. 2017) formed an additional, well-defined subcluster within ST2.

Nine (thirteen percent) of the 67 *E. coli*-positive individuals were positive for more than one sequence variant. Eight individuals had two distinct consensus sequences, while one individual generated three consensus sequences (HS0440A, HS0440B, and HS0440C; phylogenetic analyses of rDNA sequences identified the HS0440A sequence as *E. coli* ST1, whereas HS0440B and HS0440C sequences clustered with *E. coli* ST2) (Fig 1). Of note, only one of the nine individuals had two variants from within the same ST (ST2); the rest had two STs represented. The latter individuals were all from South America, and all had ST2 as one of the subtypes (ST1 + ST2 were seen in seven individuals and ST2 + ST3 were seen in one individual).

### *Entamoeba hartmanni*

For *E. hartmanni*, 23 consensus sequences were generated from 19 positive individuals. Two consensus sequences could be identified in each of four individuals. The entire dataset comprised 43 sequences, including reference sequences. Three sequence clusters could be observed: one major cluster comprised most of the sequences (n =17), whereas two minor clusters contained two and four consensus sequences, respectively (Fig 2). The major cluster was characterised by conspicuous heterogeneity, with only two of the 17 independent sequences sharing 100% identity. The two minor clades appeared more homogenous. For comparison, a phylogenetic analysis of *Entamoeba hartmanni* sequences retrieved from the NCBI database and covering the 5' end of the *SSU rRNA* gene is provided in Suppl. Fig. 2. Contrary to the situation for *E. coli*, no link between subtype and geographical origin could be observed.

## DISCUSSION

### *Entamoeba coli*

In general, molecular diversity characterisation of parasites is useful not only in basic research settings, but also in epidemiological surveillance, including for identification of lineages that differ in prevalence among healthy and diseased individuals and for identification of transmission patterns. *Entamoeba coli* is probably one of the most common eukaryotic intestinal endobionts of humans. With a cosmopolitan distribution, surveys of gastrointestinal parasites regularly report double-figure positivity rates; *E. coli* colonisation rates of up to 44.4% have been reported in some populations in Indonesia (Cross et al. 1975, Matsumura et al. 2019). Likewise, for *E. hartmanni*, colonisation rates of up to 31.3% have been reported recently (Matsumura et al. 2019).

Remarkable cryptic genetic diversity has been identified in quite a few species of intestinal endobionts, including species of *Entamoeba* (Stensvold et al. 2011, Stensvold et al. 2020, Stensvold et al. 2021b, Jacob et al. 2016, Stensvold et al. 2018, Stensvold et al. 2010). Morphologically identical endobionts can exhibit within-species genetic divergence of up to at least 30% across the entire *SSU rRNA* gene. It should be noted that no morphology data were available for any of the samples included in this study, so any potential subtype-associated differences in the cyst size of *E. coli* could not be investigated; bi- or tri-modal distributions of cyst size have been reported on more than one occasion (Dobell 1919, Matthews 1919).

The genetic diversity within *E. coli* was estimated previously to be at least 13% (Stensvold et al. 2011). The present study confirmed the genetic homogeneity of ST1 and the genetic heterogeneity of ST2 previously reported (Stensvold et al. 2011). The vast majority (85%) of the ST2 sequences in the present study were generated from faecal samples collected in South America; only two were from samples collected in Africa, and one sequence was from a sample collected in Europe. ST1 sequences, on the other hand, exhibited no particular geographical bias.

The bootstrap support for including the subcluster including reference sequence S2702 (FR686364) within ST2 is relatively low. In our previous study (Stensvold et al. 2011), this sequence was designated as being part of ST1, but also with low bootstrap support. Interestingly, the sequence was from a Nigerian woman, while the few new sequences identified from this group of *E. coli* have their origin in South America. Three other sequences in the NCBI Database also belong to this group with 99%-100% similarity: MW026736, a sequence from an individual in Brazil reported by Calegar and colleagues (Calegar et al. 2021), plus KY658178 and KY658179, which are sequences from humans in Uganda reported by Nolan and colleagues (Nolan et al. 2017) (Suppl Fig 1). Therefore, the few sequences so far identified from this group of *E. coli* have their origin in sub-Saharan Africa and South America, and no non-human host has so far been identified to harbour *E. coli* from this cluster. Based on the distinctiveness and homogeneity of the cluster, we tentatively name it *E. coli* ST3.

Only a dozen or so *E. coli* sequences from non-human primates are available in GenBank. Of these sequences, only one belongs to *E. coli* ST1 (FR686410 from *Mandrillus leucophaeus*;

Suppl Fig 1). The remainder belong to ST2, and include sequences from macaques, chimpanzees and gorillas (Suppl fig 1), and non-human primates sampled in China (KX923799–KX923802; Suppl Table 1).

The full host spectrum of *E. coli*, however, remains unclear. Sequence-based evidence of *E. coli* colonisation is readily available for both human and non-human primate hosts. However, the identity of cysts resembling *E. coli* in non-primate hosts, such as pigs and dogs, remains more obscure. Sequence data are not available, and microscopy data (including images of cysts) are also typically not available or are ambiguous. A couple of related sequences from rodents are available in the NCBI nucleotide database. The sequence FN396613 was named '*Entamoeba muris*', but our phylogenetic analysis shows that this sequence is *E. coli* ST2 (Fig 1). That sequence was reported as having been generated from a faecal DNA sample from a rat in Madrid, Spain. We previously found *E. coli* ST2 in a long-tailed chinchilla (*Chinchilla lanigera*), a rodent species native to Chile (Stensvold et al. 2011); the sequence is identical to the rat sequence mentioned above. Another sequence, FN396614, was named *Entamoeba* sp. PGF-2009 and originated from a sample from a lesser rhea, a bird native to South America; this sequence also forms part of *E. coli* ST2 (Suppl Fig 2). The combined data on *E. coli* ST2 available to date indicate that ST2 may be commonly found in various—and genetically quite different—hosts in South America. It remains unclear whether (partial) cryptic host specificity exists for any of the ST2 subclades.

Nine of the 67 *E. coli*-positive individuals had more than one sequence type, and eight of these nine were positive for both ST1 and ST2. This appears to represent an excess of mixed subtype samples given the prevalence of *E. coli* ST1 and ST2 in the South American sample set. However, given the relatively small numbers, more sampling will be needed to confirm this observation.

In conclusion, the present study confirms the large within-species genetic diversity of *E. coli* based on a dataset comprising samples from three continents, with evidence of one new subtype and indications of an uneven geographical distribution of subtypes.

### *Entamoeba hartmanni*

After aligning the *E. hartmanni* sequences for each sample, a few within-sample single nucleotide polymorphisms (SNPs) could be identified in most, making the manual selection of sequence reads for consensus sequences difficult. This issue was not observed for the *E. coli* sequences. Previous studies on genetic diversity in *E. hartmanni* relied on Sanger sequencing (Stensvold et al. 2011), where such SNPs might not have been evident. This means that the sequences included in the proposed ST1 (Fig 2) might not be as different as one might think from looking at the phylogenetic tree.

The tree clearly shows three clusters of *E. hartmanni*. In the previous analysis *of E. hartmanni* SSU rDNA sequences in 2011, such clustering was not visible, and we then concluded that *E. hartmanni* was a single robust species, in contrast to *E. coli*. However, in the 2011 study, the sequence region of study covered the first 600 bp (5' end) of the gene, and so the sequenced region was different to the one analysed in the present study. This observation highlights the relevance of obtaining as near-complete SSU rDNA sequences as possible for studies that deal with genetic diversity. For independent confirmation of the proposed subtypes, and for both *E. coli* and *E. hartmanni*, it could prove informative to sequence one or more house-keeping genes from strains representing the various clades identified to date.

In the present study, *E. hartmanni* also appears to be a species complex based on the topology of the tree and the relatively high bootstrap values for each cluster. There is no clear geographical pattern to the distribution of sequences across the three clusters, but the sample size is still small for this species, limiting our ability to make epidemiological inferences for *E. hartmanni.*

Recently, we found evidence of *E. hartmanni* in pigs sampled in Denmark (Stensvold et al. 2021b). In that study, this species was identified in 10 pigs, and all positive pigs shared the same genetic variant, which was identical to human ST1 sequence FR686375 (included in Fig 2). Thus, pigs could potentially be a reservoir for *E. hartmanni* infections in humans. Interestingly, in contrast to the human samples, the pig samples, which were also analysed by metabarcoding, did not exhibit evidence of the SNPs mentioned above.

In our supplementary analysis of *E. hartmanni*-sequences covering the 5'-end of the *SSU rRNA* gene (Suppl Fig 2), the clustering that was evident in Fig 2 is absent. This might indicate that the 5'-end of the *SSU rRNA* gene does not hold sufficient diversity to show this within *E. hartmanni.*

A complete list of the *E. hartmanni*-specific sequences in the NCBI database to date is provided in Suppl Table 2.

## Conclusion

In the present study, we confirm the existence of *E. coli* ST1 and ST2 and find evidence of an additional subtype, ST3. For *E. hartmanni*, our data suggest the existence of three subtypes, whereas previous analyses, using other regions of the *SSU rRNA* gene, failed to find genetically distinct clusters. For *E. coli,* ST1 appears to be mainly anthroponotic and to have a cosmopolitan distribution, whereas ST2 appears to be limited mostly to the southern hemisphere and with less strict host specificity. More sampling, however, might reveal cryptic host specificity within ST2. For *E. coli* ST3 and the *E. hartmanni* subtypes, data are still too limited to make conclusions on host specificity and geographic distribution.

## REFERENCES

Ascuna-Durand, K., Salazar-Sanchez, R. S., Castillo-Neyra, R. & Ballon-Echegaray, J. 2020. Relative frequency of *Blastocystis* subtypes 1, 2, and 3 in urban and periurban human populations of Arequipa, Peru. *Trop Med Infect Dis*., 5:178.

Calegar, D. A., Monteiro, K. J. L., Bacelar, P. A. A., Evangelista, B. B. C., Almeida, M. M., Dos Santos, J. P., Boia, M. N., Coronato-Nunes, B., Jaeger, L. H. & Carvalho-Costa, F. A. 2021. Epidemiology, species composition and genetic diversity of tetra- and octonucleated *Entamoeba* spp. in different Brazilian biomes. *Parasit Vectors*., 14:160.

Campos Filho, P. C., Barros, L. M., Campos, J. O., Braga, V. B., Cazorla, I. M., Albuquerque, G. R. & Carvalho, S. M. 2008. [Zoonotic parasites in dog feces at public squares in the municipality of Itabuna, Bahia, Brazil]. *Rev Bras Parasitol Vet*., 17:206-209.

Cembranelli, S. B., Souto, F. O., Ferreira-Paim, K., Richinho, T. T., Nunes, P. L., Nascentes, G. A., Ferreira, T. B., Correia, D. & Lages-Silva, E. 2013. First evidence of genetic intraspecific variability and occurrence of *Entamoeba gingivalis* in HIV(+)/AIDS. *PLoS One*, 8:e82864.

Chihi, A., O'Brien Andersen, L., Aoun, K., Bouratbine, A. & Stensvold, C. R. 2022. Amplicon-based next-generation sequencing of eukaryotic nuclear ribosomal genes (metabarcoding) for the detection of single-celled parasites in human faecal samples. *Parasite Epidemiol Control*, 17:e00242.

Chihi, A., Stensvold, C. R., Ben-Abda, I., Ben-Romdhane, R., Aoun, K., Siala, E. & Bouratbine, A. 2019. Development and evaluation of molecular tools for detecting and differentiating intestinal amoebae in healthy individuals. *Parasitology*, 146:821-827.

Cross, J. H., Clarke, M. D., Durfee, P. T., Irving, G. S., Taylor, J., Partono, F., Joesoef, A., Hudojo & Oemijati, S. 1975. Parasitology survey and seroepidemiology of amoebiasis in South Kalimantan (Borneo), Indonesia. *Southeast Asian J Trop Med Public Health*, 6:52-60.

Dobell, C. 1919. The Amoebae Living in Man. A Zoological Monograph. John Bale, Sons & Danielson, London, England.

Garcia, G., Ramos, F., Maldonado, J., Fernandez, A., Yanez, J., Hernandez, L. & Gaytan, P. 2018. Prevalence of two *Entamoeba gingivalis* ST1 and ST2-kamaktli subtypes in the human oral cavity under various conditions. *Parasitol Res.*, 117:2941-2948.

Hartmeyer, G. N., Stensvold, C. R., Fabricius, T., Marmolin, E. S., Hoegh, S. V., Nielsen, H. V., Kemp, M. & Vestergaard, L. S. 2019. *Plasmodium cynomolgi* as cause of malaria in tourist to Southeast Asia, 2018. *Emerg Infect Dis.*, 25:1936-1939.

Holmgaard, D. B., Barnadas, C., Mirbarati, S. H., Andersen, L. O., Nielsen, H. V. & Stensvold, C. R. 2020. Detection and identification of *Acanthamoeba* and other non-viral causes of infectious keratitis in corneal scrapings by real-time PCR and next-generation sequencing-based 16S-18S gene analysis. *J Clin Microbiol.*, 59: e02224-20.

Jacob, A. S., Busby, E. J., Levy, A. D., Komm, N. & Clark, C. G. 2016. Expanding the *Entamoeba* universe: New hosts yield novel ribosomal lineages. *J Eukaryot Microbiol.*, 63:69-78.

Jirků-Pomajbíková, K., Čepička, I., Kalousová, B., Jirků, M., Stewart, F., Levecke, B., Modrý, D., Piel, A. K. & Petrželková, K. J. 2016. Molecular identification of *Entamoeba* species in savanna woodland chimpanzees (*Pan troglodytes schweinfurthii*). *Parasitology*, 143:741-748.

Kobayashi, S., Suzuki, J. & Takeuchi, T. 2009. Establishment of a continuous culture system for *Entamoeba muris* and analysis of the small subunit rRNA gene. *Parasite*, 16:135-139.

Krogsgaard, L. R., Andersen, L. O., Johannesen, T. B., Engsbro, A. L., Stensvold, C. R., Nielsen, H. V. & Bytzer, P. 2018. Characteristics of the bacterial microbiome in association with common intestinal parasites in irritable bowel syndrome. *Clin Transl Gastroenterol.*, 9:161.

Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.*, 33:1870-1874.

Matsumura, T., Hendarto, J., Mizuno, T., Syafruddin, D., Yoshikawa, H., Matsubayashi, M., Nishimura, T. & Tokoro, M. 2019. Possible pathogenicity of commensal *Entamoeba hartmanni* revealed by molecular screening of healthy school children in Indonesia. *Trop Med Health.*, 47:7.

Matthews, J. R. 1919. A mensurative study of the cysts of *Entamoeba coli*. *Ann Trop Med Parasitol.*, 12:259-272.

Nolan, M. J., Unger, M., Yeap, Y. T., Rogers, E., Millet, I., Harman, K., Fox, M., Kalema-Zikusoka, G. & Blake, D. P. 2017. Molecular characterisation of protist parasites in human-habituated mountain gorillas (*Gorilla beringei beringei*), humans and livestock, from Bwindi impenetrable National Park, Uganda. *Parasit Vectors*, 10:340.

Pinilla, J. C., Morales, E. & Munoz, A. A. F. 2021. A survey for potentially zoonotic parasites in backyard pigs in the Bucaramanga metropolitan area, Northeast Colombia. *Vet World*, 14:372-379.

Poulsen, C. S. & Stensvold, C. R. 2016. Systematic review on *Endolimax nana*: A less well studied intestinal ameba. *Trop Parasitol.*, 6:8-29.

Stensvold, C. R., Lebbad, M. & Clark, C. G. 2012. Last of the human protists: The phylogeny and genetic diversity of *Iodamoeba*. *Mol Biol Evol.*, 29:39-42.

Stensvold, C. R., Lebbad, M. & Clark, C. G. 2010. Genetic characterisation of uninucleated cyst-producing *Entamoeba* spp. from ruminants. *Int J Parasitol.*, 40:775-778.

Stensvold, C. R., Winiecka-Krusnell, J., Lier, T. & Lebbad, M. 2018. Evaluation of a PCR method for detection of *Entamoeba polecki*, with an overview of its molecular epidemiology. *J Clin Microbiol,* 56: e00154-18.

Stensvold, C. R., Nielsen, M., Baraka, V., Lood, R., Fuursted, K. & Nielsen, H. V. 2021a. *Entamoeba gingivalis*: Epidemiology, genetic diversity and association with oral microbiota signatures in North Eastern Tanzania. *J Oral Microbiol.*, 13:1924598.

Stensvold, C. R., Lebbad, M., Hansen, A., Beser, J., Belkessa, S., O'Brien Andersen, L. & Clark, C. G. 2020. Differentiation of *Blastocystis* and parasitic archamoebids encountered in untreated wastewater samples by amplicon-based next-generation sequencing. *Parasite Epidemiol Control*, 9:e00131.

Stensvold, C. R., Lebbad, M., Victory, E. L., Verweij, J. J., Tannich, E., Alfellani, M., Legarraga, P. & Clark, C. G. 2011. Increased sampling reveals novel lineages of *Entamoeba*: Consequences of genetic diversity and host specificity for taxonomy and molecular detection. *Protist*, 162:525-541.

Stensvold, C. R., Jirku-Pomajbikova, K., Tams, K. W., Jokelainen, P., Berg, R., Marving, E., Petersen, R. F., Andersen, L. O., Angen, O. & Nielsen, H. V. 2021b. Parasitic intestinal protists of zoonotic relevance detected in pigs by metabarcoding and real-time PCR. *Microorganisms*, 9:1189.

Stensvold, C. R., Sorland, B. A., Berg, R., Andersen, L. O., van der Giezen, M., Bowtell, J. L., El-Badry, A. A., Belkessa, S., Kurt, O. & Nielsen, H. V. 2022. Stool microbiota diversity analysis of *Blastocystis*-positive and *Blastocystis*-negative individuals. *Microorganisms*, 10:326.

Verweij, J. J., Polderman, A. M. & Clark, C. G. 2001. Genetic variation among human isolates of uninucleated cyst-producing *Entamoeba* species. *J Clin Microbiol,* 39:1644-1646.

## FIGURE LEGENDS

**Fig 1.** Phylogenetic analysis of *Entamoeba coli*-specific sequences generated in the present study with reference *E. coli* sequences from the NCBI nucleotide database and outgroup sequences. The two previously identified subtypes (ST1, ST2) were clearly identified, as was the proposed third subtype (ST3). The Neighbor-Joining method was used. Evolutionary distances were computed using the Kimura 2-parameter method. The analysis involved 88 nucleotide sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 484 positions in the final dataset. The region covered corresponded to the middle part of the *SSU rRNA* gene (as opposed to the sequences included in Suppl Fig 1, which reflected the 5'-end of the gene). Only bootstrap values >70 are shown. Sequences generated in the present study are highlighted in boldface and were all from humans. Reference sequences are indicated in non-boldface type with information on host origin. Abbreviations used in parentheses for sample origin as follows: A, Africa; E, Europe; SA, South America.

**Fig 2.** Phylogenetic analysis of *Entamoeba hartmanni*-specific sequences generated in this study with reference *E. hartmanni* sequences from the NCBI nucleotide database and outgroup sequences. Three subtypes were identified. The Neighbor-Joining method was used. Evolutionary distances were computed using the Kimura 2-parameter method. The analysis involved 43 nucleotide sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There was a total of 411 positions in the final dataset. The

region covered corresponded to the middle part of the *SSU rRNA* gene (as opposed to the sequences included in Suppl Fig 2, which reflected the 5'-end of the gene). Only bootstrap values >70 are shown. Sequences generated in the present study are highlighted in boldface and were all from humans. Abbreviations used in parentheses for sample origin as follows: A, Africa; E, Europe; SA, South America.

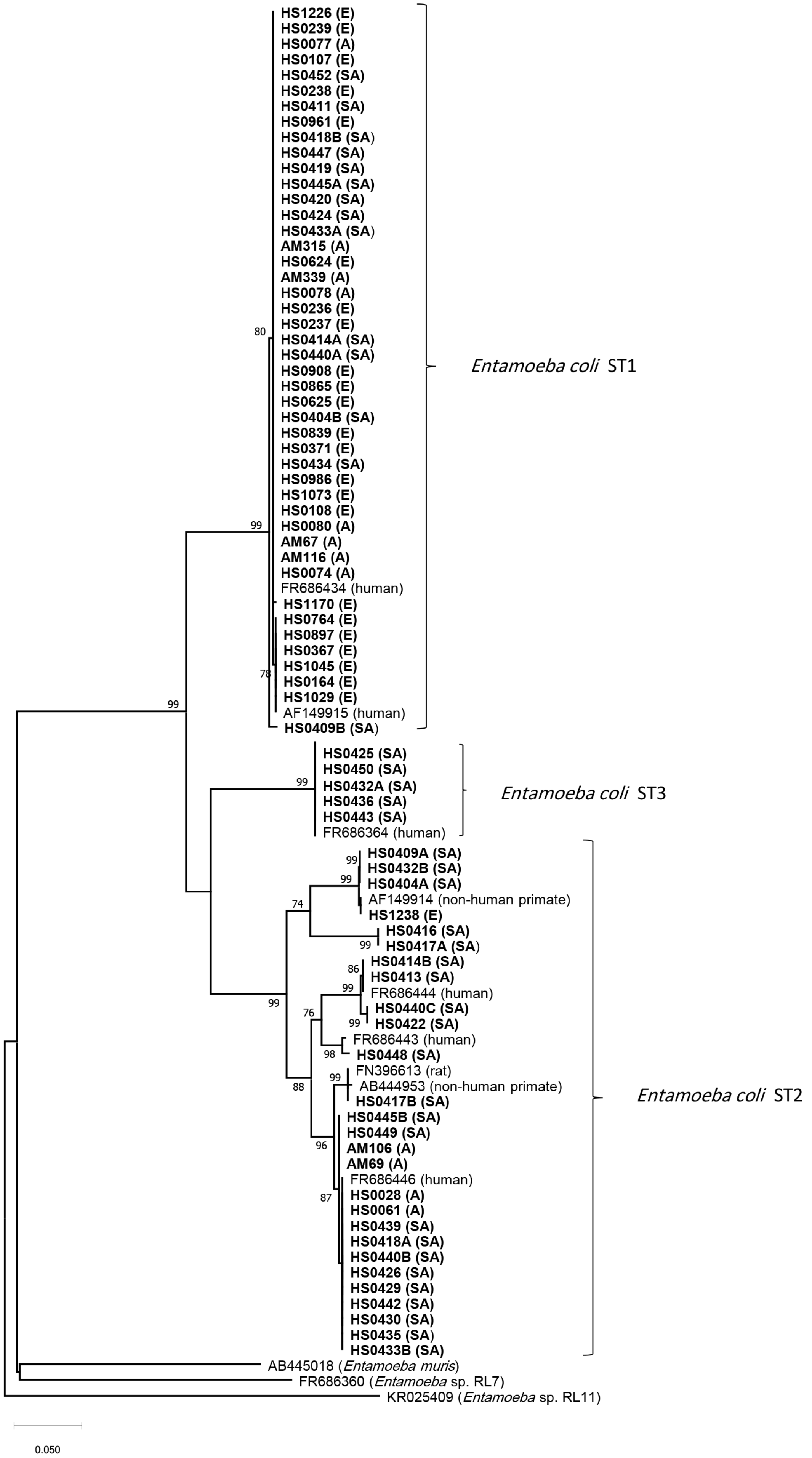
## SUPPORTING INFORMATION

**Fig S1.** Intraspecies variability of *Entamoeba coli* based on 104 DNA sequences representing the 5'-end of the small subunit rRNA gene retrieved from the NCBI database. The *E. coli*-specific sequences obtained from non-human hosts are indicated in bold-face type. Non-human *E. coli* host names are indicated in parentheses after the NCBI Database ID accession number. Three subtypes are acknowledged. The Neighbor-Joining method was used. Evolutionary distances were computed using the Kimura 2-parameter method. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were 244 positions in the final dataset. Only bootstrap values >70 are shown.

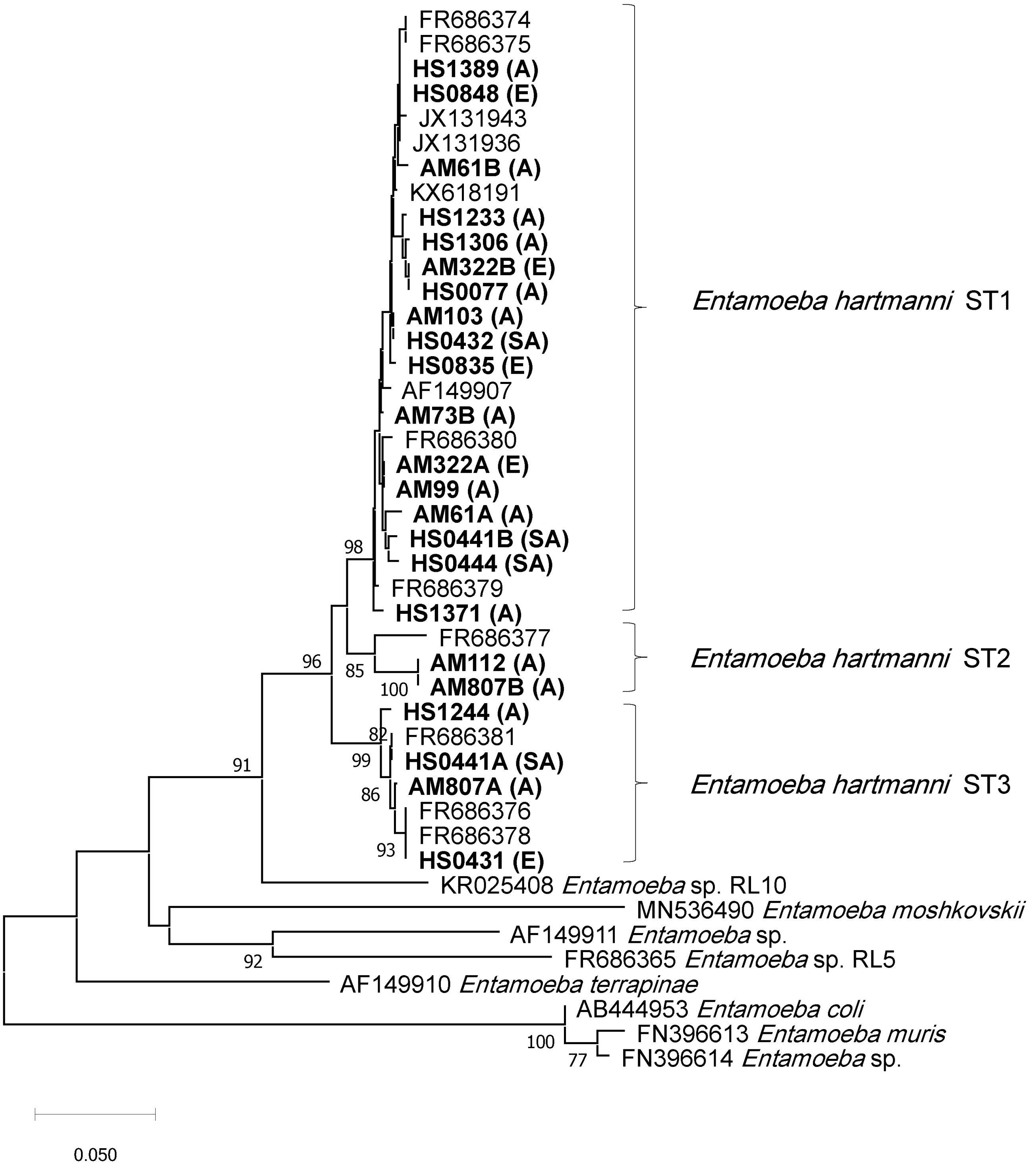
**Fig S2.** Intraspecies variability of *Entamoeba hartmanni* based on the 5'-end of the small subunit rRNA gene. The *E. hartmanni*-specific sequences obtained from non-human hosts are indicated in bold-face type. Non-human *E. hartmanni* host names are indicated in parentheses after the NCBI Database ID accession number. The region covered corresponded to the 5'-end of the gene (as opposed to the sequences included in Fig 2, which reflected the middle part of the *SSU rRNA* gene). Based on this part of the gene, the existence of the three subtypes could not be observed The Neighbor-Joining method was used. Evolutionary distances were computed using the Kimura 2-parameter method. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were 502 positions in the final dataset. Only bootstrap values >70 are shown.

**Table S1.** Sequences, STs, and hosts of *Entamoeba coli* identified to date.

**Table S2.** Sequences, STs, and hosts of *Entamoeba hartmanni* identified to date.



JEU_12949_Figure -01-TIFF 300 dpi.tiff



JEU_12949_Figure -02-TIFF 300 dpi.tiff