microorganisms

MDPI

Article

RNA Sequencing Reveals Widespread Transcription of Natural Antisense RNAs in *Entamoeba* Species

Damien Mornico [1,*], Chung-Chau Hon [2,3,*], Mikael Koutero [4], Christian Weber [2,3], Jean-Yves Coppée [4], C Graham Clark [5], Marie-Agnes Dillies [1,4] and Nancy Guillen [2,3,6,*]

1 Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, 75015 Paris, France; marie-agnes.dillies@pasteur.fr
2 Institut Pasteur, Unité Biologie Cellulaire du Parasitisme, 75015 Paris, France; christian.weber@pasteur.fr
3 Institut National de la Santé et de la Recherche Médicale, INSERM U786, 75015 Paris, France
4 Institut Pasteur, Plate-Forme Transcriptome et Epigénome, 75015 Paris, France; mikael@koutero.me (M.K.); jean-yves.coppee@pasteur.fr (J.-Y.C.)
5 London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; Graham.Clark@lshtm.ac.uk
6 Centre National de la Recherche Scientifique, CNRS ERL9195, 75015 Paris, France
* Correspondence: damien.mornico@pasteur.fr (D.M.); chungchau.hon@riken.jp (C.-C.H.); nguillen@pasteur.fr (N.G.)

Abstract: *Entamoeba* is a genus of Amoebozoa that includes the intestine-colonizing pathogenic species *Entamoeba histolytica*. To understand the basis of gene regulation in *E. histolytica* from an evolutionary perspective, we have profiled the transcriptomes of its closely related species *E. dispar*, *E. moshkovskii* and *E. invadens*. Genome-wide identification of transcription start sites (TSS) and polyadenylation sites (PAS) revealed the similarities and differences of their gene regulatory sequences. In particular, we found the widespread initiation of antisense transcription from within the gene coding sequences is a common feature among all *Entamoeba* species. Interestingly, we observed the enrichment of antisense transcription in genes involved in several processes that are common to species infecting the human intestine, e.g., the metabolism of phospholipids. These results suggest a potentially conserved and compact gene regulatory system in *Entamoeba*.

Keywords: parasite; genomics; transcriptomics; antisense RNA

Citation: Mornico, D.; Hon, C.-C.; Koutero, M.; Weber, C.; Coppée, J.-Y.; Clark, C.G.; Dillies, M.-A.; Guillen, N. RNA Sequencing Reveals Widespread Transcription of Natural Antisense RNAs in *Entamoeba* Species. *Microorganisms* **2022**, *10*, 396. https://doi.org/10.3390/microorganisms10020396

Academic Editors: Christen Rune Stensvold and Anastasios D. Tsaousis

Received: 7 December 2021
Accepted: 5 February 2022
Published: 8 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Entamoeba organisms are endobiotic amoebae colonizing species of animals. The persistence of at least seven species of *Entamoeba* depends on their ability to infect humans, mainly in the intestinal tract, where they divide and encyst [1]. There has been a renewed interest in commensal intestinal amoebae because, being members of the human eukaryome, these microorganisms can be an important component in the establishment and functioning of intestinal homeostasis [2,3]. One of these species is *E. histolytica*, the etiological agent of amebiasis. As a pathogenic parasite, *E. histolytica* varies between the commensal status seen in 90% of infected people and the virulent status that leads to intestinal invasion in the remaining 10% [4]. Other species of *Entamoeba* are intestinal commensals and include *E. dispar* which has a life cycle similar to that of *E. histolytica* [1]. *E. histolytica* and *E. dispar* together account for nearly 82% of *Entamoeba* infections in humans. Infections with *E. dispar* seems to be ~10 times more common than *E. histolytica* [5]. The remaining *Entamoeba* infections correspond to other non-pathogenic species, *E. coli* (1.98%), *E. hartmanni* (0.96%), *E. polecki* (0.04%) and *E. gingivalis* (4.6%) [5]. Another species related to *E. histolytica* is *E. moshkovskii*, a free-living amoeba commonly found in sewage and known to infect humans, although its pathogenicity is not precisely determined [6,7]. Overall, most *Entamoeba* infections in humans seems to be caused by commensal species and these species feed on and evolve with *Enterobacteriaceae*. Another

widely studied species is *E. invadens*, which is pathogenic in a wide range of reptiles [8,9]. *E. invadens* is used as a model for the studying the formation of amoebic cysts, due to the possibility of completing its cell cycle under laboratory conditions.

The genomes of *E. histolytica* strain HM1:IMSS [10–12]; *E. dispar* strain SAW760 [13], *E. moshkovskii* strain Laredo [14] and *E. invadens* strain IP-1 [15,16] have been sequenced. All these genomes are AT-rich, although *E. invadens* and *E. moshkovskii* genomes have unusually biased GC distributions [14]. *E. histolytica* and *E. dispar* are closely related species, bearing 95% and 85% of nucleotide identity in their genic and intergenic regions, respectively [17,18]. The genome of *E. moshkovskii* is slightly larger than that of *E. histolytica*, and the phylogenetic relationship between diverse isolates of *E. moshkovskii* indicates that their most recent common ancestor is 500 times more ancient than the ancestor of isolates from *E. histolytica* [14]. The parasite of reptiles, *E. invadens*, is readily distinguishable from the other human infecting species, sharing only 74% and 62% of nucleotide identity in their genic and intergenic regions, respectively [15,16]. Despite transposable elements being abundant in all four genomes [19,20], there are clear distinctions among species: While DNA transposons dominate in *E. moshkovskii* and *E. invadens*, retrotransposons dominate in *E. histolytica* and *E. dispar* [21]. These transposons are often located near genes encoding tRNA, which are organized in tandem arrays [22,23]. *E. histolytica* has the most compact genome among the four species, with ~8300 coding genes in 20 Mb, compared to the *E. invadens* genome which bears 11,549 coding genes in 40 Mb [14]. The gene lengths of *E. histolytica* and *E. invadens* are similar while the intergenic regions in *E. invadens* are longer than those in *E. histolytica* (408 bp vs. 282 bp) [16]. A total of 4704 gene families comprising 21,741 genes are shared among all four species [14].

Transcriptomic analyses enabled the identification of gene promoter sequences in E. invadens and *E. histolytica* [24–28], as well as the characterization of small RNAs regulating encystation of *E. invadens* and gene silencing in *E. histolytica* [29–31]. More recent analyses using RNA sequencing (RNA-Seq), essentially performed in *E. histolytica*, enabled genome-wide identification of transcription start sites (TSS) [32] and polyadenylation (polyA) sites (PAS) [33,34], as well as regulatory RNA such as miRNAs [35], small RNAs [29] and long non-coding RNAs [36]. More recently, natural antisense RNAs (NATs) [32], which are non-coding RNAs transcribed from the opposite strand of a gene, have been described in *E. histolytica*. NAT transcription is abundant as roughly 25% of genes generate NATs in all environmental conditions tested [32]. In addition, genomic sequences for TSS and PAS are similar for sense and antisense transcription, which indicates a novel transcriptional regulatory scenario probably derived from the compactness of the *E. histolytica* genome [32]. These characteristics of RNA biogenesis in the related species *E. dispar*, *E. moshkovskii* and *E. invadens* have not yet been completely described.

In this study, we aimed to study the conservation of gene regulatory sequences (i.e., motifs at TSS and PAS) and NAT transcription between *E. histolytica*, *E. dispar*, *E. moshkovskii* and *E. invadens*. We analysed similarities and differences in transcriptional genomic determinants for mRNA and NATs and concluded that TSS for NATs occurs on the opposite strand within the coding sequence, although variations exist between the species. NATs precisely initiate at the stop codon in *E. histolytica* and *E. dispar* but appear to be dispersed inside the intergenic spacers in the two other species. Comparing the transcriptomic profiles of genes targeted by NATs in *E. histolytica* to *E. dispar* and *E. moshkovskii*, we discovered significant biological processes common to amoebic species infecting the human intestine, with lipid metabolism activities as the most important enriched functions, whereas the comparison with *E. invadens* highlighted vesicular trafficking and regulation of RNA transcription. Common gene products and functions associated with NATs identify robust candidates for monitoring NAT biogenesis in all four species of Entamoebidae.

2. Materials and Methods

2.1. Entamoeba Species and In Vitro Culture

Entamoeba histolytica strain HM-1: IMSS was isolated in 1967 from a biopsy of a rectal ulcer from an adult male with amoebic dysentery, Mexico City, Mexico. *Entamoeba dispar* strain SAW760 was isolated in 1979 from an asymptomatic adult male at the London School of Hygiene and Tropical Medicine, UK. The Laredo strain of *Entamoeba moshkovskii* was isolated from a resident of Laredo, TX, USA, who showed symptoms of diarrhoea. *Entamoeba invadens* strain IP-1 was isolated in Canada from a natural infection of a painted turtle *Chrysemys picta*. The HM-1: IMSS strain was a gift from Professor Ruy Perez Tamayo and Dr Alfonso Olivos (UNAM, Mexico). The three other strains are from the collection of Professor Graham Clark. All the strains were cultured in axenic media whose compositions have been described previously [37]. *E. histolytica* strain HM-1: IMSS was cultured in TYI-S-33 medium at 37 °C; *E. moshkovskii* strain Laredo was cultured in LYI-S-2 medium at room temperature; *E. dispar* strain SAW760 was cultured in LYI-S-2 medium at 37 °C and *E. invadens* strain IP-1 was cultured in TYI-S-33 medium at 25 °C.

2.2. RNA Preparation

Total RNA was extracted from approximately 1×10^6 trophozoites (with each culture performed in triplicate) using Trizol (Invitrogen, Thermo Fisher Scientific, Waltham, Massachusetts, U.S.A, ref. 15596026). The poly(A) fraction was purified from 10 to 100 μg of total RNA using Dynabeads according to the manufacturer's instructions (Thermo Fisher Scientific ref. 28152103011150). Small RNAs were purified from 10 μg of total RNA loaded on a denaturing 15% TBE- Urea gel (BioRad, Marnes-La-Coquette, France, ref. 456-6053). More specifically, following ZR-small-RNA Ladder (Zymo Research, Tustin, California, U.S.A, ref. R1090) size indications, a fragment of gel corresponding to the 15–35 nt region was excised with a scalpel and RNAs were extracted using the ZR small-RNA PAGE Recovery Kit (Zymo Research, ref. R1070) following manufacturer's recommendations. Small RNAs were then treated with Tobacco Acid Pyrophosphatase (Epicentre Biotechnologies, Madison, Wisconsin, U.S.A, ref. T19100) for 1 h and purified by RNA Clean and Concentrator-5 (Zymo Research, ref. R1015) following the manufacturer's instructions.

2.3. Library Construction, Sequencing and Read Processing

First, libraries were constructed using TruSeq Small RNA Sample Prep Kit (Illumina, San Diego, California, U.S.A, ref. RS-200-0012) following the manufacturer's instructions. Then, all the libraries were purified on a 5% TBE gel (BioRad, ref. 456-5013) and were quantified by Bioanalyzer DNA High Sensitivity Chips (Agilent Technologies, Santa Clara, California, U.S.A, ref. 5065-4626). Sequencing was performed on HiSeq-2000 (Illumina) in a multiplexed 51 + 7 nucleotide single-end read using a TruSeq SR Cluster kit v3 cBot HS (Illumina, ref. GD-401-3002) and a TruSeq SBS kit v3 HS 50 cycles (Illumina, ref. FC-401-3002). Finally, FASTQ sequence files were generated using CASAVA 1.8.2 (Illumina) and adapter sequences were trimmed using AlienTrimmer [38] (v. 0.4.0). Reads shorter than 20 nucleotides were discarded.

2.4. Reference Genomes

The reference genome sequences and annotations of *E. histolytica* (20.80 Mb, 1496 contigs and 8201 genes), *E. dispar* (22.96 Mb, 3312 contigs and 8744 genes), *E. moshkovskii* (25.25 Mb, 1147 contigs and 12,260 genes) and *E. invadens* (40.88 Mb, 1144 contigs and 11,997 genes) were downloaded from AmoebaDB v46 (https://www.amoebadb.org/common/downloads/release-46, accessed on 5 January 2022).

2.5. Data Repositories

The data are available in the SRA database (https://www.ncbi.nlm.nih.gov/sra/) under accession number PRJNA781395.

2.6. De Novo Assembly of Transcriptomes

Libraries within each species (Table S1) were merged for mapping and de novo assembly. First, reads were mapped to the corresponding reference genomes of each species using STAR v2.5.0 [39] with a maximum intron length of 900 (—alignIntronMax), maximum mismatches of 2 (—outFilterMismatchNmax), maximum multi-mapping locations of 10 (—outFilterMultimapNmax) and minimum overhang of 25 (—alignSJoverhangMin) for spliced alignments. Mapped reads were split into plus and minus strand-sets with Samtools v. 1.13 [40]. Then, de novo assembly of transcript fragments (i.e., contigs) were performed on each strand-set using Trinity [41], with a kmer size of 15 nt (—KMER_SIZE 15), a maximum intron size of 900 nt (—genome_guided_max_intron 900), a minimum coverage of 20 (—genome_guided_min_coverage 20) and a minimum length of 150 nt (—min_contig_length 100). Lastly, resulting transcript contigs were mapped on the corresponding reference genome using Gmap [42] with a maximum intron length of 900 nt (-K 900). Transcript contigs mapped to the opposite strands of coding genes (with a minimum 10% contig length, detected using featureCounts software [43] with —fracOverlap 0.1) were defined as "NAT fragments". Genes were defined as "NAT genes", when at least 1 NAT fragment mapped. In order to specify parts of genes preferentially covered by NAT fragments, we split the genes in 5 equal regions and counting was performed on each region as well.

2.7. Analyses of Small RNAs

The small RNA data were processed as previously described [44]. Briefly, deduplicated reads were removed using fqCleanER (https://gitlab.pasteur.fr/GIPhy/fqCleanER, accessed on 10 August 2021). The remaining unique reads were then mapped to the corresponding reference genomes using Bowtie v. 0.12.7 [45] with maximum 1 end-to-end mismatch (-v 1) and all mapped locations were reported (—all). Genes with ≥20 small RNA reads were defined as sRNA targeted genes.

2.8. TSS Identification and Annotation

The reads were aligned to the different reference genomes, using Bowtie v. 0.12.7 [45] with the following parameters: maximum of 2 mismatches were allowed (-*n* 2) and reads mapped to multiple locations (-m 50) were reported only once (-k 1). The alignments produced were sorted and indexed with SAMTools [40]. Coverage graphs representing the numbers of mapped reads per nucleotide were generated based on the sorted reads using BEDTools [40,46], focusing on 5′ end position (−5). On each coverage, an upper quartile normalization [47] was performed and a minimum coverage of 4 was imposed. TSS candidates within 10 nts from each other were then clustered together in transcription initiation clusters and the position of the strongest coverage was defined as the peak. Each TSS was then classified as a gene TSS (gTSS), an internal TSS (iTSS), an antisense TSS (asTSS), or an orphan (oTSS) if it could not be assigned to any of the previous classes [48]. A TSS was classified as a gTSS if it was located ≤100 bp upstream of a gene and as an asTSS if it was located within the 200 bp surrounding the stop codons. The TSS with the strongest expression values (maximum peak height) among gTSS of a gene was classified as primary (pTSS). iTSS were located within an annotated gene on the sense strand.

2.9. PAS Identification and Annotation

First, reads with a stretch of five or more 'A's at their ends (or 'T's at their beginning) were selected for this analysis, as they potentially contain mRNA poly(A) tails. Redundant

reads were removed and stretches of As at the ends were trimmed. Remaining reads with a minimum length of 18 nt were then mapped on the reference genome using Bowtie [45] with following parameters: -*n* 2 -k 1 -m 50 -l 30. To avoid false positives due to sequencing errors, reads with low quality (<20) around PAS (5 nt upstream and downstream) were removed from the set. To discriminate real poly(A) tracks of true polyadenylation from poly(A) tracks of internal homopolymeric stretches on the mRNAs, false positives were discarded if they met the following criteria: (i) reads with ≥8 nt within 10 nt immediately upstream of the PAS are A's, (ii) mapping with ≥5 'A's immediately downstream of the PAS.

PAS candidates within 12 nts from each other were then clustered together in PAS clusters and the position of the strongest coverage was defined as the peak. PAS with fewer than 2 reads of coverage at the peak were removed.

2.10. Motif Enrichment

The sequences immediately upstream and downstream of the gTSS, and asTSS (100 nt on each side), as well as the PAS of mRNA and NAT were used to scan for conserved motifs using DREME [49]. The immediate upstream or downstream sequences were thus used as the positive sets, and the farther upstream (at position −200) or downstream (at position +150) sequences of the same length were used as the negative sets. To visually investigate the positional enrichment of these discovered motifs surrounding the polyadenylation sites, the total occurrence of these motifs was searched along the sequences surrounding (300 nt) the poly(A) sites

2.11. Orthologs, Core and Pan-Genome Identification

The four species' proteomes were used to compute orthologous groups of genes among all strains with OrthoFinder v. 2.3.8 [50], with the Blast sequence comparison option (-S blast) and default MCL inflation parameter (-I 1.5). Orthologous groups with genes present in each species as a unique copy were identified as "core-genome". Genes involved in synteny groups have been identified in AmoebaDB (https://amoebadb.org, accessed on 5 January 2022).

2.12. Differential Expression Analysis

Firstly, reads of each replicate were mapped to the reference genome of *E. histolytica*, using STAR v. 2.5.0 [39]) with a maximum intron length of 900 (—alignIntronMax), 2 mismatches maximum (—outFilterMismatchNmax), 10 locations maximum for a read mapping (—outFilterMultimapNmax), minimum overhang of 25 nt for spliced alignments. Secondly, read counts for each gene in each sample were computed with featureCounts [43] separately on the direct (-s 1) and opposite strand (-s 2), allowing multi-mapping reads (-M). In order to take into account length differences between orthologous genes across species, we first applied a Transcript Per Million (TPM) quantification [51]. The DESeq2 normalization was used: size factors were computed on sense counts and were then applied to both sense and antisense counting. Finally, transcript differential expressions were calculated on the merged normalized counts (sense and antisense) using DESeq2 v. 1.24.0 [52] within the SARTools pipeline v. 1.7.2 [53]. False discovery rates were corrected using the Benjamini–Hochberg procedure. The analysis was conducted in R [54] and figures were produced using the ggplot2 package [55].

2.13. eggNOG Annotations and GO Terms Enrichment

To combine gene datasets the Venn diagram approach was used (https://bioinfogp.cnb.csic.es/tools/venny/, accessed on 15 December 2021). Predicted proteins were retrieved using UniProt (https://www.uniprot.org/, accessed on 15 December 2021) using gene name and the Uniprot function. For annotation, predicted proteins of *Entamoeba* were assigned to the orthologous groups of the EggNOG database v. 5 [56] using the eggNOG-mapper. V. 2 (http://eggnog-mapper.embl.de/, accessed on 27 August 2021) according to their defined parameters (hit e-values ≤ 1e-3). One-letter abbreviations for the functional categories' correspondence are at https://www.ncbi.nlm.nih.gov/research/cog#, accessed on 27 August 2021. Further identification of protein classes and gene ontology enrichments corresponding to genes harbouring sense and antisense transcripts were performed with Amoeba DB tools: search for gene ID, data analysis for GO terms (biological processes and molecular functions) and synteny search in the setting pre-configured table, orthologs and paralogs within VEuPathDB (https://amoebadb.org/amoeba/app, accessed on 28 December 2021). For proteins related to lipid metabolism analysis we used UniProt (https://www.uniprot.org/, accessed on 28 December 2021) and STRING (https://string-db.org/ accessed on 28 December 2021) clustering methods.

3. Results

3.1. Widespread NAT in Entamoeba Species

To compare the transcriptomes between the four species of *Entamoeba*, RNA-seq was performed on polyadenylated RNAs (Table S1). De novo assembly of the RNA-seq data yielded 35,000 to 57,000 transcript fragments (i.e., contigs), with mean lengths of 251 to 403 nt, among the four species (Table S2). In total, 15,624 to 27,687 of these contigs can be mapped to the sense strand of genes in their corresponding genome, covering 83% to 95% of genes, attesting to the good representation of the global transcriptome from each species (Figure 1A and Table S3, Sheet 1). From 11% to 27% of these contigs are mapped in intergenic regions and are probably derived from unannotated or non-coding genes (Table S2). Interestingly, 6401 to 13,462 contigs, with mean lengths between 205 nt and 339 nt, can be mapped to the antisense strand of genes, covering 38% to 59% of the genes among the species, suggesting widespread NAT in *Entamoeba* (Figure 1A and Table S3, Sheet 2). Of the four species, *E. dispar* showed the greatest proportion of genes with NAT (~59%). Previous studies have shown that NATs are shorter than the corresponding gene and more likely to map to the 3′ region in *E. histolytica* [32]. After dividing the coding sequences into five equal sized regions, we investigated the coverage pattern of the mRNA and NAT contigs among the four species (Figure 1B). From 65% to 83% of genes are fully covered by mRNA contigs, homogeneously between regions, confirming the good distribution of the transcriptome. The 3′ biased mapping of NAT contigs was present in *E. histolytica*, *E. dispar* and *E. moshkovskii*, while *E. invadens* displayed a uniform pattern of coverage.

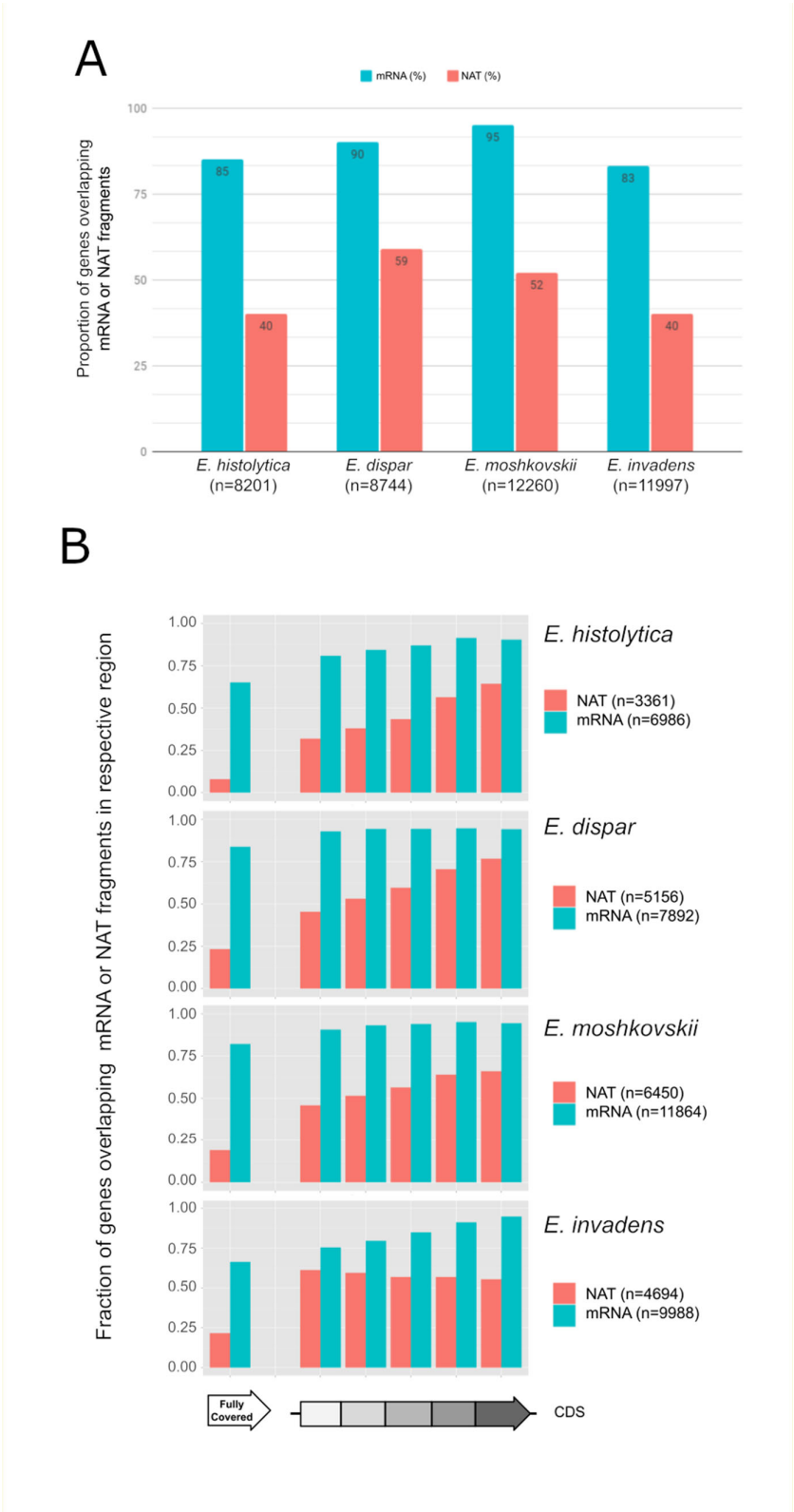


Figure 1. NATs and mRNA among the species. (**A**) Proportion of CDSs linked to mRNAs (blue) and NATs (red) in the 4 species. (**B**) Proportion of genes overlapping at least one mRNA (blue) or NAT (red) contig in the 5 different regions of their CDS on both sides in each species.

3.2. Small RNAs and NATs Are Independent Entities of Non-Coding RNAs

Previous studies demonstrated the existence of a distinct class of 27 nt small RNAs in *E. histolytica* and *E. invadens* [29,57,30,44]. Here we sought to investigate whether the biogenesis of NAT and these small RNAs are related or not. We thus purified and sequenced the small RNAs in the four *Entamoeba* species and investigated the correlation of the genome-wide distributions of small RNAs and NAT in each species. As formerly highlighted [44], two main populations of sRNA are identified in *E. histolytica* and *E. invadens*, based on their size distribution (Figure S1A), with two peaks at 27 nt and 31 nt. The same distribution is observed in *E. dispar* and *E. moshkovskii* sRNA populations. Among the four species, ~8% to ~10% of genes were found to overlap antisense small RNAs (Figure S1B and Table S4, Sheet 2), which is much lower than the genes overlapping NATs, described above and only 9 of these genes are common to all the species (according to the orthology analysis described below). In addition, we noticed a bias in the distribution of small RNAs towards the 5′ end of genes in *E. histolytica* and *E. dispar*, and a more spread-out distribution in *E. moshkovskii* and *E. invadens* (Figure S1C), in contrast to the general bias of NAT toward the 3′ end of genes. Then, we used χ^2 tests in order to further investigate associations between sRNA and NAT distributions (Table S4, Sheet 1). Independence of distributions was rejected in each species. We observed that sRNA and NATs tend to exclude each other. In the genomes we analysed, there was a higher number of NAT genes found in the absence of sRNA (>92%) than when both co-occurred. Reciprocally, a higher number of sRNA genes were found without NATs (from 55% to 71%). Co-occurrence of both NATs and sRNA (in blue) was low compared to expected values (green). We concluded that strong positive association between NATs and sRNA was unlikely. The lack of correlation between the genome-wide distributions of NAT and small RNAs suggests the biogenesis of these two classes of RNAs might be independent.

3.3. Identification of TSS for mRNA and NATs

We have recently reported that most NATs in *E. histolytica* are initiated from the 3′ end of the coding sequences (CDS), and in particular at the stop codon, which acts as the Initiator (Inr) motif at the TSS of NATs [32]. In this study, we sought to investigate this aspect in other *Entamoeba* species. Using the method previously described [32], we mapped the TSS in the genomes of all four species. Gene TSSs (gTSS), defined as the TSS closest to the start codons (–100 nt) on the sense strand, were identified in *E. histolytica* (*n* = 4213, in 3649 genes), *E. dispar* (*n* = 2878, in 2543 genes), *E. moshkovskii* (*n* = 3561, in 3221 genes) and *E. invadens* (*n* = 2032, in 1756 genes). Antisense TSSs (asTSS), defined as TSSs around the stop codon (+/–100 nt) on the antisense strand, were also mapped in *E. histolytica* (*n* = 1710, in 1197 genes), *E. dispar* (*n* = 1869, in 969 genes), *E. moshkovskii* (*n* = 3033, in 1162 genes) and *E. invadens* (*n* = 1667, in 639 genes) (Tables S5). While gTSS are strongly enriched around 10 nt upstream the start codons, the distribution of asTSS is not as well-defined among the four species (Figure 2A). Indeed, we observed a sharp peak of asTSS at the stop codon in *E. histolytica* and *E. dispar*, while the asTSS in *E. moshkovskii* and *E. invadens* showed a spread distribution at the 3′ end of the CDSs (Figure 2A). These differences could indicate different genomic sequences regulating the initiation of NATs between the two closest species, *E. histolytica* and *E. dispar*, compared to the others (Figure 2B). Indeed, the Inr motifs of gTSS and asTSS seem similar to the stop codon in *E. histolytica* and *E. dispar*; it is the case for gTSS in *E. moshkovskii and E. invadens* although bases C and G appeared more often in their gTSS. The tendency to include C and G in Inr motifs from *E. moshkovskii* and *E. invadens* is even more pronounced in the case of asTSS, precluding a clear-cut conclusion from these data concerning PolII preferences in *E. moshkovskii* and *E. invadens*.

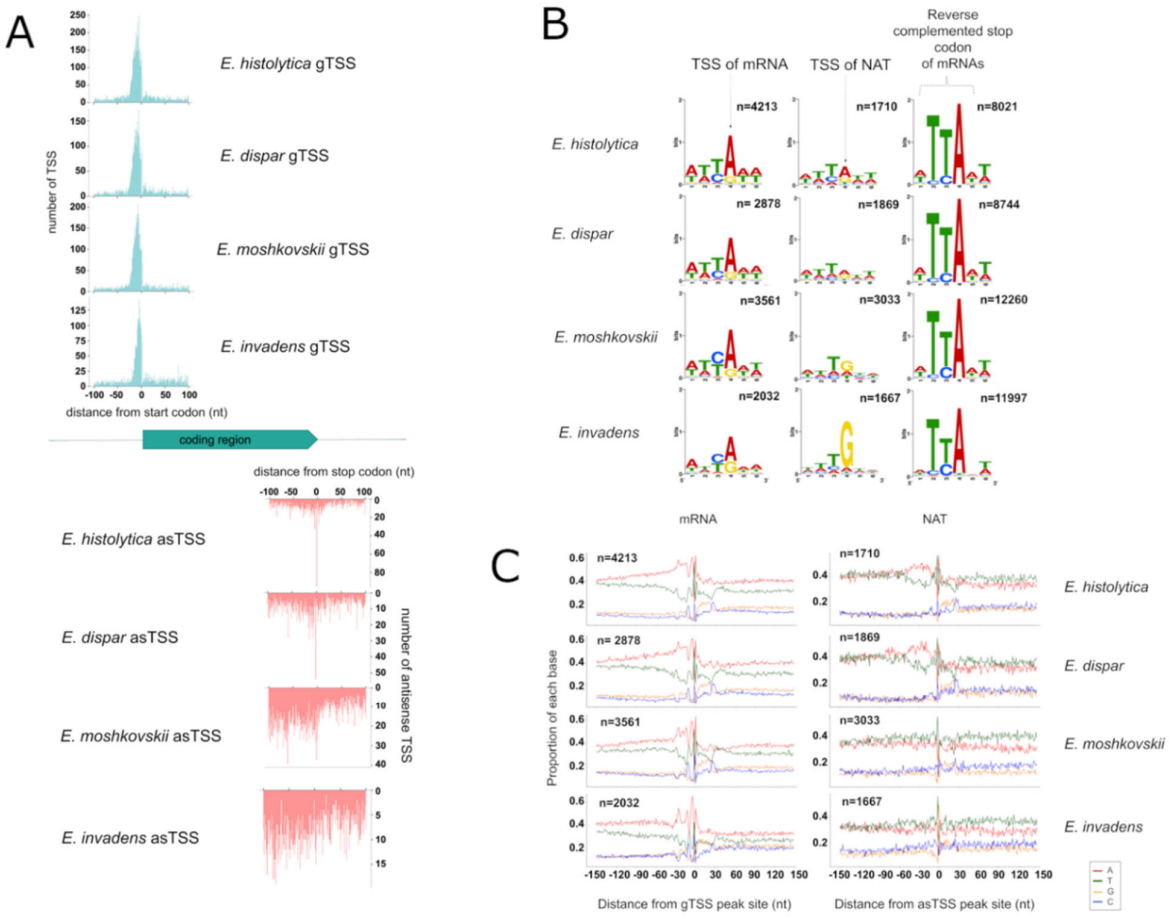


Figure 2. Transcription start sites of mRNA and NATs. (**A**) Mapping of TSS at CDS boundaries (with 100 nt apart) on both strands in the different species. (**B**) Sequences logo computed around stop codon using the entire genome CDS reverse complemented (right), asTSS (middle) and mRNA TSS (left) found in different species. (**C**) Nucleotide sequence composition around TSS of both mRNA and NAT. When several are assigned to a same gene, only the strongest TSS is considered (primary). Notice that 0 corresponds to the TSS peak identified.

Sequence compositions surrounding the TSS of mRNA (primary gTSS) and ta-NAT (primary asTSS) were compared between the four species (Figure 2C). *E. histolytica* and *E. dispar* display clear similarities in nucleotide proportion on both sense and antisense strands, including (i) an A-rich region around −80 to −20 nt, (ii) a T/A enriched region (at −30 nt) within this A-rich region (iii) a C/T enriched region around −10 nt, (iv) a YA motif around TSS (e.g., Inr in Figure 2C), and (v) a C/G enriched region at +25 nt, implying the sequence determinants to initiate mRNA and NAT transcription are essentially the same in these two species. These regions are also visible around *E. moshkovskii* and *E. invadens* gTSSs, although with slightly different proportions, but not clearly defined around asTSSs.

3.4. Identification of PAS for mRNAs and NATs

Based on the methods used in our previous study [32], we identified PAS genome-wide in the four species and defined them as mRNA PAS or NATs PAS based on their strands (Table S6). By that definition, 2640 mRNA PAS and 1935 NAT PAS were detected in *E. histolytica*, and similarly, 4835 and 4015 in *E. dispar*; 5789 and 4609 in *E. moshkovskii* and 5825 and 2584 in *E. invadens*. The median distances of mRNA PAS and NAT PAS from

the stop-codon are similar in the four species, at 20 to 24 nt and 417 to 494 nt, respectively (Figure 3A). These features are consistent with the previous results [32] and confirm that NATs are mostly shorter than mRNA in *Entamoeba* species.

Analysis of sequence composition around the PAS shows clear similarities among the four species, for both mRNA and NAT (Figure 3B), except for the U-rich region downstream of the PAS (U-rich DSE), which seems shorter in *E. invadens* (from 1 to 15 nt after PAS) than in the others (from 1 to 26 nt after PAS). The cleavage sequence element (CSE), representing the 3 nucleotides around the PAS, is also well identified in all species but slightly less significantly enriched in *E. invadens*. These observations imply that the genomic determinants of polyadenylation are not only conserved between mRNA and NAT, but also conserved among the different species, with slight variations in *E. invadens*, probably reflecting its greater evolutionary distance to the other species.

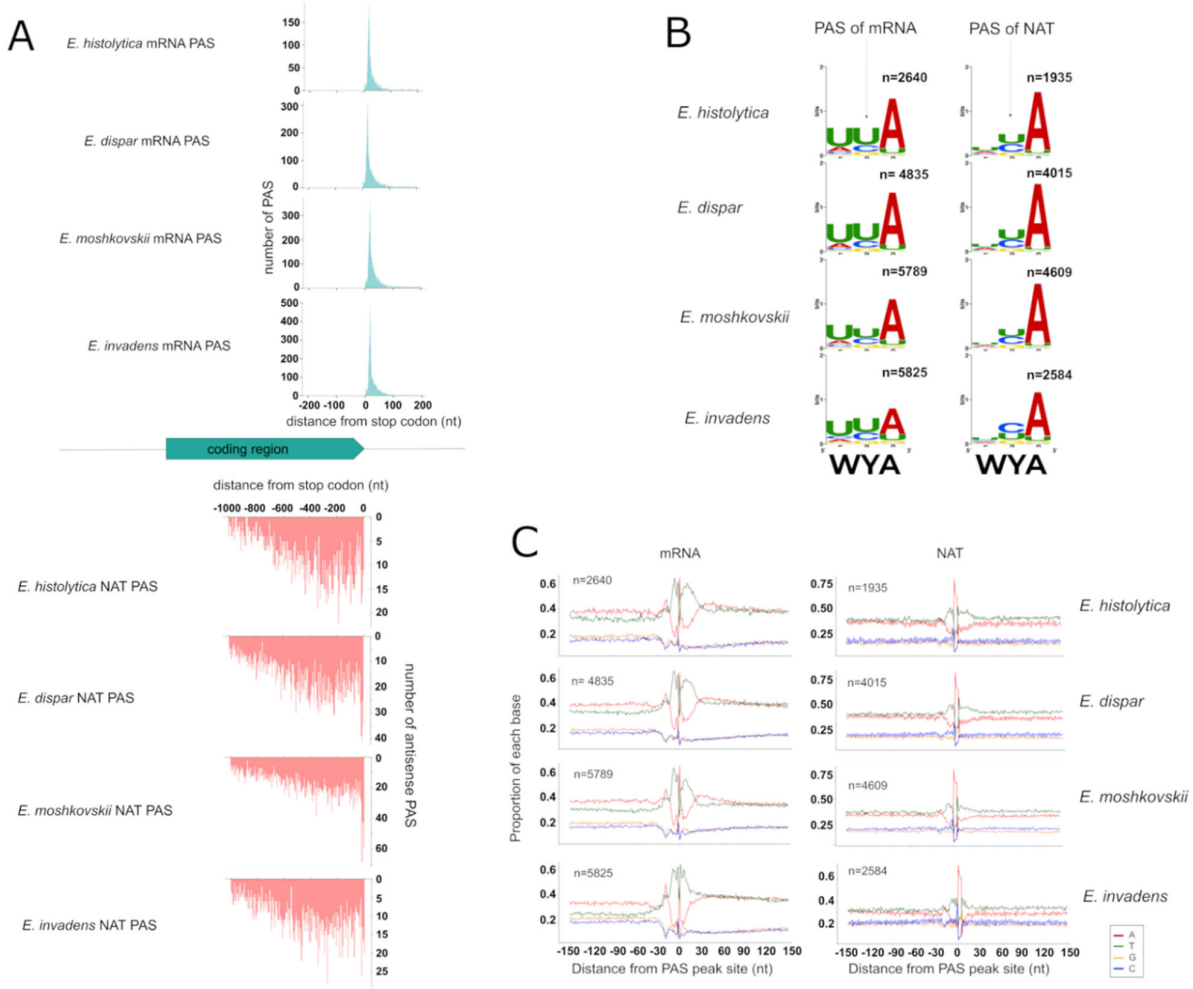


Figure 3. Poly(A) site identification and motif enrichment. (**A**) Distribution of PAS distance from stop codon (position 0) on sense (blue) and antisense DNA strand (red). (**B**) Sequence logo around the PAS of both mRNA (left) and NAT (right). (**C**) Nucleotide sequence composition around the RNA cleavage site (position 0 corresponds to 1 nt before the cleavage site) of both mRNA (left) and NAT (right).

3.5. NAT in the Core-Genome and Species-Specific Genes

To investigate the conservation of NATs, we defined groups of orthologous genes among the four species (see Methods). We identified a set of genes that are orthologous among all four species (n = 5069), and 3565 of them exist as a unique copy in each species (Figure 4A and Table S3, Sheet 3). Herein we define these 3565 genes as the core-genome, and the genes without an orthologous partner as species-specific genes in each species. Next, we compared the extent of NAT transcription in the core-genome among species. While the orthologs in *E. histolytica* and *E. invadens* showed a similar extent of NAT transcription (~46%), the extent is substantially higher in *E. moshkovskii* and *E. dispar* (65% and 72% respectively) (Figure 4B and Table S3, Sheet 4). Thus, the extent of NAT transcription appears to be higher in the core-genome than in the whole individual genomes, but variations were observed between species.

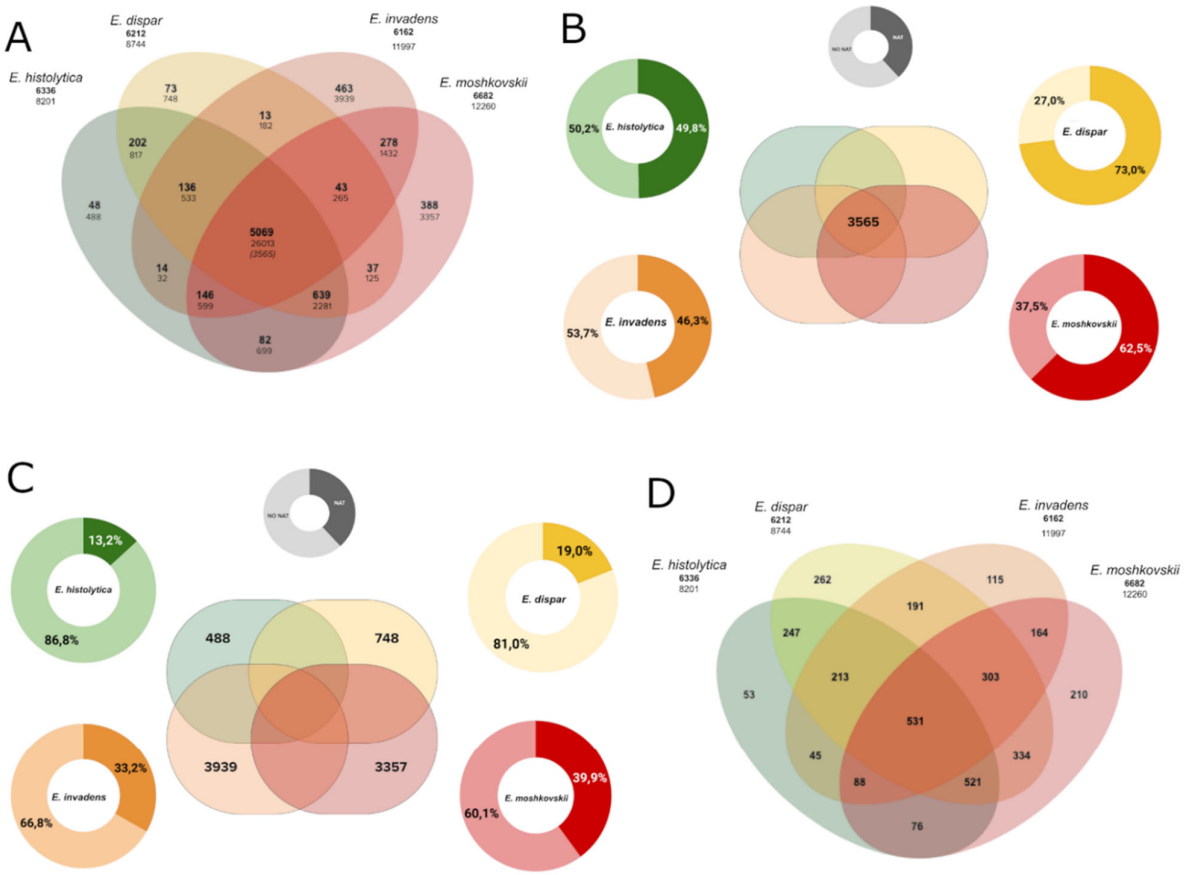


Figure 4. NATs in the core- and pan-genome. (**A**) Venn diagram of specific and orthologous genes among the 4 species. Numbers in bold represent gene families and numbers in regular font represent the genes in those families. The number in italics is the number of families with genes in a unique copy. (**B**) Proportion of NATs in the core-genome for each species. (**C**) Proportion of NAT in specific genes for each species. (**D**) Venn diagram of specific and common NATs in the core-genome.

We also compared the extent of NAT transcription in the species-specific genes. As *E. histolytica* and *E. dispar* have closely related genomes, only a few genes were identified as species-specific (488 and 748, respectively), compared to *E. moshkovskii* and *E. invadens* which harboured large numbers of species-specific genes (3357 and 3939, respectively) (Figure 4A and Table S3, Sheet 7). Among these species-specific genes, *E. histolytica* and *E. dispar* displayed a similar extent of NAT transcription (12% and 19% of genes respectively), compared to the greater extents in *E. invadens* and *E. moshkovskii* (33% and 39%, respectively) (Figure 4C and Table S3, Sheet 8). We noted that these numbers are smaller than the global antisense expression recorded for whole genomes, and variation between genomes is not comparable. *E. dispar*, for instance, showed the smallest extent of NAT transcription in species-specific genes, while it has the largest extent of NAT transcription in the core-genome.

3.6. Functional Annotations of NAT-Associated Genes Conserved in All Species

About 14% of genes in the core-genome (n = 531) are associated with NAT in all four species (hereinafter named the coreNATcore, Figure 4D and Table S3, Sheet 5) and more than half of these (n = 271) are annotated as "unknown functions" in AmoebaDB. To investigate the functional enrichment of genes in the coreNATcore, we annotated their potential functions based on orthology mapping using eggNOG-mapper [56] and 360 of them can be mapped to a cluster of orthologous groups (COG) with functional annotations (Table S7, Sheet 1). These include regulation of RNA biosynthesis and transcription (COG categories K, A and J, n = 63); signal transduction mechanisms and trafficking (COG categories T and U, n = 72); posttranslational modifications, protein turnover and chaperones (COG categories O, n = 30); DNA replication, repair, chromatin structure and chromosomes partition (COG categories L, D and B, n = 15); transport and metabolism of carbohydrates, lipids, coenzymes, amino acids and ions (COG categories G, F, I, H, E and P, n = 46); cytoskeleton (COG categories Z and CZ, n = 6); and other COG categories (n = 24). In addition, 98 proteins have a recognizable domain but with unknown function (COG category S, n = 98). We also annotated the genes in the core-genome using eggNOG-mapper (Table S7, Sheet 2), and 2425 of the 3551 genes can be mapped to a COG. In the core-genome, 20% to 34% of genes within the COG categories N (cell motility), Y (nuclear structure), F (transport and metabolism of nucleotides), H (coenzyme), D (cell cycle), K (transcription), E (amino acids) were associated with NATs (Figure 5) and the COG category W (extracellular structures) was not found in genes associated with NATs. The other mapped functional categories (less than 20%) are summarized in Figure 5.

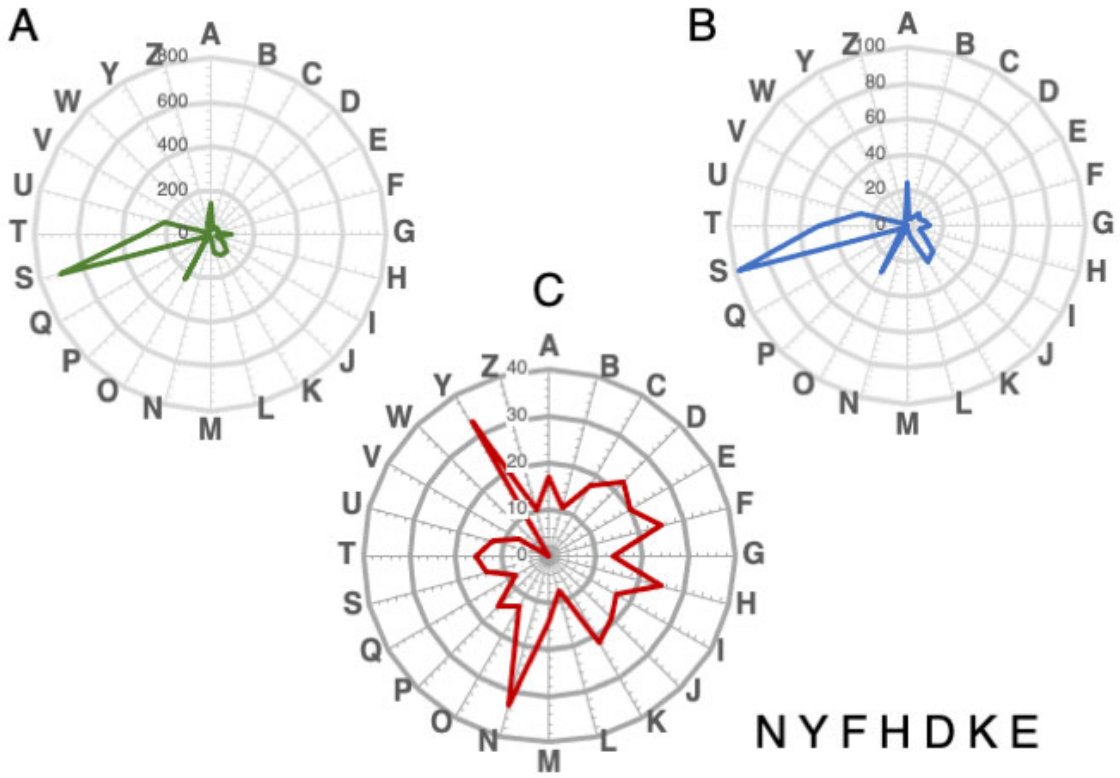


Figure 5. Functional categories of genes targeted by NATs in the core-genome and in the coreNATcore**.** Predicted proteins of *Entamoeba* from the core-genome and from genes in the coreNATcore were assigned to clusters of orthologous groups of proteins (COG) in the EggNOG database. (**A**) Categories of orthologous groups found in the core-genome. (**B**) Categories of orthologous groups found in the coreNATcore dataset. (**C**) percentage of gene-presenting NATs in relation to the core-genome. One-letter abbreviations for the functional categories correspondence at COG–NCBI (https://www.ncbi.nlm.nih.gov/research/cog, accessed on 27 August 2021) are: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell division and chromosome partitioning; E, amino acid metabolism and transport; F, nucleotide metabolism and transport; G, carbohydrate metabolism and transport; H, coenzyme metabolism; I, lipid metabolism; J, translation, including ribosome structure and biogenesis; K, transcription; L, replication, recombination and repair; M, cell wall structure and biogenesis and outer membrane; N, motility; O, molecular chaperones and related functions; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general functional prediction only; S, no functional prediction; T, signal transduction; U, intracellular trafficking secretion and vesicular transport; V, defence mechanisms; W, extracellular structures; Y, nuclear structure; Z, cytoskeleton.

Previously, we defined a set of NAT-associated genes (n = 452) that are common among *E. histolytica* strains under various environmental conditions (defined as *E. histolytica* core NAT) [32], including invasion of the human colon. From those 452 genes, in this work (Table S3, Sheet 5), 98% of them (n = 444) are present in that dataset of *E. histolytica* genes presenting NATs. Next, we examined in a Venn diagram the overlap between these *E. histolytica* core NATs and the inter-species coreNATcore defined above and 86 genes were found to be present in both datasets (Table S8, Sheets 1 and 2). In the four species, there is no significant difference between the level of expression of these 86 genes and the level of expression of the other genes according to the comparison test of Wilcoxon (p-value = 0.1099, 0.1425, 0.1727 and 0.4273 for *E. histolytica*, *E. dispar*, *E. moshkovskii* and *E. invadens* respectively), (Supplementary Figure S2). They map into 57

contigs (Table S8, Sheet 3) and are not clustered in any particular region within these contigs. Examining the gene ontology (GO) terms of these 86 inter-species common genes (Table S8, Sheet 4), we highlight seven genes involved in metabolism of small molecules, including EHI_000720 which encodes a 5′-methylthioadenosine/S-adenosylhomocysteine (MTA/SAH) nucleosidase. It is involved in the S-adenosyl-L-methionine (SAM, AdoMet) cycle, which recycles S-adenosyl-L-homocysteine back to SAM, and in salvage pathways for 5′-deoxyadenosine and S-methyl-5′-thioadenosine [58]. MTA/SAH nucleosidase plays a key role in the purine salvage pathway and in recycling of methylthio groups, the encoding gene (EHI_000720) is believed to have been acquired by the *E. histolytica* genome through lateral gene transfer (LGT) from a bacterial genome. The six other genes correspond to long-chain-fatty-acid-CoA ligase EHI_029050 (catalysing β-oxidation of fatty acids); triosephosphate isomerase EHI_056480 (enzyme related to glycolysis); galactokinase EHI_094100 (catabolism of galactose); pyridoxal kinase EHI_126090 (phosphotransferase, Vitamin B6 metabolism); enolase EHI_130700 (enzyme related to glycolysis) and leucyl-tRNA synthetase EHI_161970 (ligation of L-leucine to tRNA). Apart from genes with GO term annotations, other proteins involved in known metabolic processes reinforced the above observations. For example, EHI_155520 encodes a protein carrying the C-terminal catalytic domain of glutamine synthetase involved in nitrogen assimilation. Three leucine-rich-repeat-containing proteins (EHI_023340 EHI_024640 EHI_148460), which were also acquired by the *E. histolytica* genome through LGT from bacterial genomes, are of unknown function. Another three proteins containing LIM zinc finger domains (EHI_022960, EHI_110280 and EHI_194520), several kinases and small GTPAses are expected to have signalling actions through the cytoskeleton. In addition, EHI_107280 encodes a V-type ATPase from V0 complex regulating cytoplasm *p*H, which is a homologue of EHI_078250 vacuolar protein sorting protein 26 of the amoebic retromer-like complex [59]. The Ubiquitin fusion degradation protein 1 (EHI_125920), which interacts with the nuclear localization protein 4 (Npl4 in yeast and EHI_026450 in *E. histolytica*) forms a cofactor for AAA-family ATPase CDC48 for proteasome-dependent processing of ubiquitinated ER-associated proteins [60,61]; in addition, Npl4 is involved in the transport of polyadenylated RNA from the nucleus to the cytoplasm [62]. A zinc finger protein EHI_008770 presents homology to Tristetraprolin, which binds to AU-rich elements in mRNA, leading to the removal of the poly(A) tail from the mRNA and increasing rates of mRNA decay [63]. A 5′-3′ exonuclease EHI_080270 has homology to Xrn1 exonuclease, the enzyme catalysing cytoplasmic mRNA degradation in decay pathways [64]. In summary, these 86 NAT-associated genes encode factors involved in metabolic processes, biosynthesis and turnover of RNA, as well as in vacuolar traffic and protein degradation by the proteasome that are common to all four species of *Entamoeba*.

3.7. Expression of NATs in Syntenic Blocks

The assembly and gene annotation of *Entamoeba* genomes are largely incomplete, complicating comparisons of the transcriptomes between species. To investigate the conservation of NAT between species, we examined the expression of NAT in syntenic blocks based on the syntenic annotations in AmoebaDB. In the core-genome (n = 3565), 3496 and 926 genes from *E. histolytica* are syntenic with at least a second species and with all four species, respectively (Table S3, Sheet 3). For the coreNATcore (n = 531), 524 are syntenic between at least two species and 141 genes are syntenic in all the four species (Table S9, Sheet 1). These 141 syntenic genes are distributed among 79 contigs of the *E. histolytica* genome (Table S9, Sheet 2), with 26 contigs containing two or more genes targeted by NAT and 53 contigs contain only one NAT targeted gene (Table S9, Sheet 3). In this visual contig inspection, we did not find any correlation between the proportion of gene associated with NATs, nor with their physical mapping or the number of genes on the contigs (Supplementary Figure S3). We then applied the analysis to orthologous genes (syntenic or not) presenting NATs in *E. histolytica* (1774 genes, Table S9, Sheet 5). These are in 315 contigs, 53 of which carry only one NAT-targeted gene (Table S9, Sheet 6). The

largest contig (DS57114 of 530624 bp carrying 268 genes) was then scanned. We found 74 NAT-targeted genes which are distributed along the contig; therefore, they do not form clusters. Individually located genes are numerous (n = 21) and non-overlapping pairs are also observed. The last map according to their forward or reverse transcription sense and positioned in a convergent, opposite or similar direction (Supplementary Figure S4). Overall, the data indicating that there is no evident systematic link between physical association of genes and NAT transcription in linear genomes.

3.8. Differential Gene Expression Profile of Entamoeba Histolytica Antisense RNA Transcription

We have previously profiled the gene expression changes in *E. histolytica* upon environmental change, including parasite invasion of the human colon [65]. In line with our interest in the discovery of amoebic factors related to intestinal infections, we took advantage of interspecies transcriptomic data generated in this study (three replicates from each species) to identify transcriptomic signatures specific to *E. histolytica.* Performing differential expression analyses on the core-genome (see Methods), we identified significantly differentially expressed *E. histolytica* mRNAs or NATs (fold changes > 3, fdr < 0.05) compared to another species (Figure 6 and Table S10). Among the three species, *E. invadens* showed the largest extent of differential expression, with 1544 mRNAs (856 up-regulated and 682 down-regulated) and 1688 NATs (1240 up-regulated and 448 down-regulated) differentially expressed. *E. moshkovskii* showed a moderate extent of differential expression, with 1095 mRNAs (587 up-regulated and 508 down-regulated) and 1580 NATs (1282 up-regulated and 298 down-regulated) differentially expressed. *E. dispar*, which is phylogenetically closest to *E. histolytica* [19], showed the smallest extent of differential expression, with 750 mRNAs (455 up-regulated and 295 down-regulated) and 1046 NATs (756 up-regulated and 290 down-regulated).

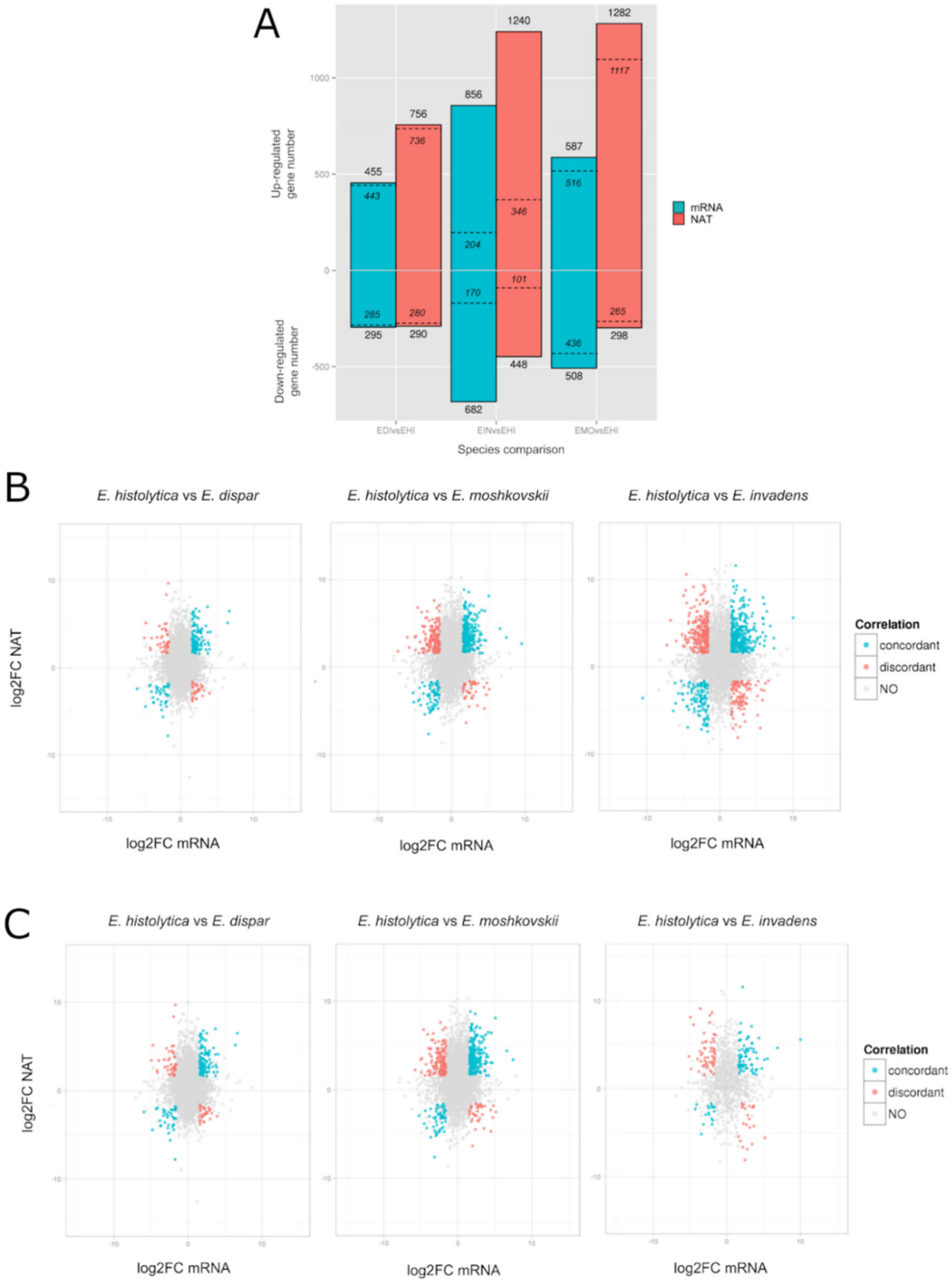


Figure 6. Differential expression of *E. histolytica* genes in comparison with the three other species. (**A**) Number of up-regulated and down-regulated mRNA and NATs, detailed in Table S10. Dashed lines represent the corresponding number of regulated genes in synteny between the two species. (**B**) The fold change in gene expression (FC) was compared by plotting the log2FC of mRNA (x axis) versus log2FC of NAT (y axis) for each gene having at least one NAT contig identified, for each comparison. The colour of points illustrates the differential expression type: none or unidirectional (grey), both concordant (blue), both discordant (red). (**C**) Same illustration keeping only genes in synteny between each pair of species.

We also explored whether the inter-species differential expression of *E. histolytica* mRNAs is related to NATs. Interestingly, we found the up-regulated, but not the down-regulated, *E. histolytica* mRNAs are significantly associated with the presence of NATs (χ^2 tests, $p < 0.05$, odds ratio > 1, Table S11A–C in all three species). To further investigate this phenomenon, we examined the associations between the mRNA differential expression and NATs differential expression. Surprisingly, we found significant associations (χ^2 tests, $p < 0.05$, odds ratio > 1, Table S11D–H in all three species), independent of the direction of differential expression (i.e., up- or down-regulation) of the mRNAs or NATs. Further examining the fold changes of mRNAs and their corresponding NATs across species (Figure 6B), we can divide the mRNA/NAT pairs into "concordant" and "discordant", based on the relative direction of their differential expression [66]. In general, we found more concordant mRNA/NAT pairs than discordant pairs. Specifically, we found concordant pairs that are up- (n = 113, 240 and 320) or down-regulated (n = 39, 65 and 113) for both mRNA and NATs expressions, and discordant pairs with up-regulated mRNA and down-regulated NAT (n = 30, 38 and 99) or down-regulated mRNA and up-regulated NAT (n = 41, 149 and 222). The significant associations of the inter-species differential expression mRNAs and NATs suggest the potential regulatory functions of NATs, or the potential linkage between biogenesis of NATs and mRNAs.

3.9. Functional Annotations of E. histolytica Syntenic Genes with Differentially Expressed NATs

Finally, we explored functional significance of *E. histolytica* NAT-associated genes with differentially expressed NATs between species. To avoid over-sophisticated interpretations, here we considered only syntenic genes. In the core-genome (n = 3565), there are 3450, 3108 and 911 *E. histolytica* genes are syntenic with *E. dispar*, *E. moshkovskii* and *E. invadens*, respectively (Table S3, Sheet 3). Examining the significantly differentially expressed NATs (fold changes > 3, fdr < 0.05) in these syntenic *E. histolytica* genes, *E. moshkovskii* showed the largest number with 1382 NATs (1117 up-regulated and 265 down-regulated), followed by *E. dispar* with 1016 NATs (736 up-regulated and 280 down-regulated) and then *E. invadens* with 447 NATs (346 up-regulated and 101 down-regulated) (Figure 6A and Tables S1–S14). Among these, there are 102 and 58 NATs that are up-regulated and down-regulated across all comparisons. Next, we examined the annotated functions of these genes with differentially expressed NATs based on Gene Ontology (GO) terms.

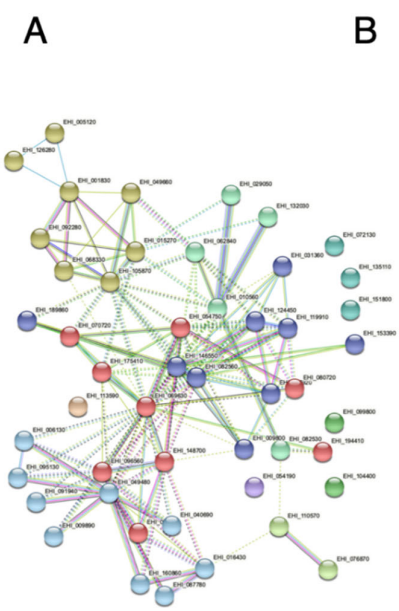
In comparison to *E. dispar*, among the 736 *E. histolytica* genes with up-regulated NATs, the significantly enriched processes are linked to lipid metabolism with 27 genes including the phosphatases necessary for the dephosphorylation of phospholipids (e.g., myotubularin); 31 genes are involved in small GTPases signal regulation, 13 genes are necessary for ubiquitin-dependent protein catabolic process, 3 genes are involved in DNA damage response, 3 genes are linked to protein export from the nucleus and 2 genes are implicated in RNA decapping including methylguanosine-cap decapping, which inactivates translation initiation and promotes 5′-to-3′ decay of mRNA (Table S12A). The 280 *E. histolytica* genes with down-regulated NATs can be annotated with 82 GO terms (Table S12B). Notably, they include proteins for cell adhesion, e.g., the Gal-GalNAc lectin subunit Igl2 (EHI_18300) and another under-described lectin (EHI_108490). DNA metabolic processes were also highly represented by several GO Terms with six genes involved in chromatin modification, DNA damage responses through DNA repair and DNA methyltransferase activity.

In comparison with *E. moshkovskii*, among the 1117 *E. histolytica* genes with up-regulated NATs, 423 of them are also up-regulated NATs in comparison with *E. dispar*. These genes shared 10 GO terms including lipid metabolism (n = 36) and nuclear export of proteins (n = 3). In addition, diverse global metabolic pathways were found (n = 94); as were response to stimuli (n = 92), where small GTPases signal pathways are represented by 29 genes, and DNA damage (n = 23, e.g., double-strand break repair MRE11, DNA

mismatch repair protein PMS1 and ligases for DNA repair), while eight genes were involved in the metabolism of RNA (for example, the activity of aminoacyl t-RNA ligase) (Table S13A). The 265 *E. histolytica* genes with down-regulated NATs shared 20 GO terms with genes also with down-regulated NATs in comparison to *E. dispar* (Table S13B), including histone acetylation (n = 2) and protein dephosphorylation (n = 2). Specific to *E. moshkovskii* NAT down regulation there was supplemental protein dephosphorylation (n = 5), chromatin organization (n = 4), secretion by cells (n = 3), the transcription factor TFIID subunit (n = 1) and cAMP signalling involving adenylate cyclase activity (n = 1).

In comparison with the relatively distant *E. invadens*, the 346 *E. histolytica* genes with up-regulated NATs can be annotated with 43 GO terms (Table S14A). These include proteolysis, essentially ubiquitination and proteasome activity (n = 17); protein modification (n = 11), including endocytosis (e.g., clathrin adaptors AP2 and AP3) or protein translocation (e.g., SRPa, Vsp16, Vsp26, coatomer β, coatomer g, SR particle). Organonitrogen compound catabolic process (n = 10) which, in addition to ubiquitination process and proteasome degradation, also showed MTA/SHA nucleosidase, the transcription elongation factor B and RAD23 which plays a central role both in proteasomal degradation of misfolded proteins and DNA repair; two genes involved in gene silencing including the transcription regulator silent information regulator 2 (EHI_151300) and the NAD-dependent deacetylase (EHI_120280); two genes implicated in positive regulation of transcription: the transcription elongation factor SPT4 (EHI_148350) and the mediator of RNA polymerase II transcription subunit 7, (EHI_170300); four genes encoding factors necessary for tRNA modification including queuine tRNA-ribosyltransferase; the Elp1 subunit of the elongation protein complex, required for tRNA modification; and tRNA pseudouridine synthase. The other GO terms common to the other species includes eight GO terms shared with *E. dispar*, five shared with *E. moshkovskii* and two shared with both amoebae species containing few genes (e.g., phosphatidylinositol-mediated signalling, nucleoside catabolism, glycosyl compound catabolism). For the 101 down-regulated NATs the enrichment of GO terms identifies few genes. These are related to protein ubiquitination (n = 3), protein maturation (n = 2) and vesicular retrograde transport (n = 2) (Table S14B). Shared with *E. moshkovskii* is the presence of adenylate cyclase but there is no GO term uniquely in common with *E. dispar*.

In summary, the profile of *E. histolytica* compared to those of *E. dispar* and *E. moshkovskii* rank lipid metabolism activities as the most important enriched function specific to the species infecting humans (e.g., lipases, kinases, phosphatases and GPI biogenesis), which networking by STRING clearly distinguished the most significant gene clusters (Figure 7). Vesicular trafficking, protein degradation and RNA metabolism were found in the comparison of the four species, but regulation of RNA transcription is most pronounced in the comparison with *E. invadens*. DNA repair functions are highly specific in the comparison with *E. moshkovskii*. These findings offer new research alternatives for a better understanding of the intestinal colonization by *Entamoeba*.



cluster	cluster color	gene count	Gene ID	protein description
1	Red	9	EHI_054750	Phosphatidate cytidylyltransferase, putative
1	Red	9	EHI_069630	CDP-diacylglycerol--inositol3-phosphatidyltransferase, putative
1	Red	9	EHI_070120	Myotubularin, putative
1	Red	9	EHI_070720	L-myo-inositol-1-phosphate synthase
1	Red	9	EHI_080720	Aminoalcoholphosphotransferase ; Belongs to the CDP-alcohol phosphatidyltransferase class-I family
1	Red	9	EHI_096560	Phosphatidylinositol 3-kinase, putative ; Belongs to the PI3/PI4-kinase family
1	Red	9	EHI_148700	Phosphatidylinositol 4-kinase, putative ; Belongs to the PI3/PI4-kinase family
1	Red	9	EHI_175410	Inositol-1-monophosphatase; Myo-inositol monophosphatase, putative
1	Red	9	EHI_194410	Isopentenyl-diphosphate delta-isomerase, putative
2	Brown	1	EHI_113590	Uncharacterized protein (Alcohol acetyltransferase domain)
3	Dark Golden Rod	8	EHI_001830	Uncharacterized protein (PIG-T, glycosylphosphatidylinositol (GPI) trans-amidase domain)
3	Dark Golden Rod	8	EHI_005120	Uncharacterized protein (Phosphatidylinositol-glycan biosynthesis class S, PIG-S domain)
3	Dark Golden Rod	8	EHI_015270	N-acetylglucosaminyl-phosphatidylinositol de-N-acetylase, putative
3	Dark Golden Rod	8	EHI_049660	Uncharacterized protein (N-acetylglucosaminyl-phosphatidylinositol de-N-acetylase, PIG-L domain)
3	Dark Golden Rod	8	EHI_068330	Mannosyltransferase involved in glycosylphosphatidylinositol- anchor biosynthesis
3	Dark Golden Rod	8	EHI_092280	GPI-anchor transamidase, putative
3	Dark Golden Rod	8	EHI_105870	Phosphatidylinositol N-acetylglucosaminyltransferase subunit A, putative
3	Dark Golden Rod	8	EHI_126280	Uncharacterized protein (GPI transamidase subunit PIG-U domain)
4	Green Yellow	2	EHI_076870	3-oxo 5 alpha-steroid 4-dehydrogenase domain-contai; Steroid 5-alpha reductase, putative
4	Green Yellow	2	EHI_110570	Protein tyrosine phosphatase, putative
5	Green 2	2	EHI_099800	Lipase, putative
5	Green 2	2	EHI_104400	Lipase, putative
6	Green	5	EHI_010560	Acyl-CoA synthetase, putative
6	Green	5	EHI_029050	Long-chain-fatty-acid--CoA ligase, putative
6	Green	5	EHI_062840	UDP-N-acetylglucosamine--dolichyl-phosphate N-acetylglucosaminephosphotransferase, putative
6	Green	5	EHI_082530	Acetyl-coA carboxylase, putative
6	Green	5	EHI_132030	Long-chain-fatty-acid--CoA ligase, putative
7	Blue	3	EHI_072130	Lipase, putative
7	Blue	3	EHI_135110	Uncharacterized protein (Ubiquitin and 3-oxo-5-alpha-steroid 4-dehydrogenase domain)
7	Blue	3	EHI_151800	Ganglioside gm2 activator protein, putative
8	Light Sky Blue	9	EHI_006130	Phosphatidylinositol 3-kinase, putative ; Belongs to the PI3/PI4-kinase family
8	Light Sky Blue	9	EHI_009890	Phosphatidylinositol-4,5-bisphosphate 3-kinase, putative ; Belongs to the PI3/PI4-kinase family
8	Light Sky Blue	9	EHI_016430	Myotubularin ; Belongs to the protein-tyrosine phosphatase family. Non-receptor class myotubularin subfamily
8	Light Sky Blue	9	EHI_040690	Phosphatidylinositol 3-kinase, putative ; Belongs to the PI3/PI4-kinase family
8	Light Sky Blue	9	EHI_045480	Phosphoinositide kinase, putative
8	Light Sky Blue	9	EHI_087780	Myotubularin ; Belongs to the protein-tyrosine phosphatase family. Non-receptor class myotubularin subfamily
8	Light Sky Blue	9	EHI_091940	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit; Belongs to the PI3/PI4-kinase family
8	Light Sky Blue	9	EHI_095130	Phosphatidylinositol 3-kinase, putative ; Belongs to the PI3/PI4-kinase family
8	Light Sky Blue	9	EHI_160860	Inositol polyphosphate 5-phosphatase, putative
9	Medium Blue	9	EHI_009800	Phosphatidylserine synthase, putative
9	Medium Blue	9	EHI_031360	Lecithin:cholesterol acyltransferase domain-containing protein
9	Medium Blue	9	EHI_067920	CDP-alcohol phosphatidyltransferase ; Belongs to the CDP-alcohol phosphatidyltransferase class-I family
9	Medium Blue	9	EHI_082560	Phospholipase D, putative
9	Medium Blue	9	EHI_119910	Lipid phosphate phosphatase, putative
9	Medium Blue	9	EHI_124450	Lipid phosphate phosphatase, putative
9	Medium Blue	9	EHI_146550	Phospholipase D, putative
9	Medium Blue	9	EHI_153390	Phospholipase, patatin family protein
9	Medium Blue	9	EHI_189860	Glycerophosphoryl diester phosphodiesterase, putative
10	Purple	1	EHI_054190	Lipase, putative
	common to EDI and EMO			

Figure 7. Lipids metabolism as a major target of antisense transcription in *Entamoeba* infecting humans. The test was carried out with the sum of lists of syntenic genes involved in lipid metabolism in *E. histolytica* in comparison with *E. dispar* or with *E. moshkovskii*, and which were targeted by NAT (up-regulated). A single dataset was submitted to the STRING program for gene functions clustering (K-means clustering methods). (**A**) Note the metabolism of phospholipids (red and light sky blue), biosynthesis of GPI (dark gold rod) and lipid degradation (medium blue) as the most representative among the 10 clusters analysed. (**B**) Genes found in both *E. dispar* and *E. moshkovskii* are underlined in blue, while those not underlined are the genes found in only one of these parasites.

4. Discussion

Seven species of *Entamoeba* colonize humans with different frequency; three of them are currently candidates for study because they represent more than 90% of cases of amoebic infection, namely *E. histolytica*, *E. dispar* and *E. moshkovskii*. In this work, we described and compared the transcriptomes of these three human infecting species, and a reptile infecting species, *E. invadens*, aiming to understand their phenotypic differences. All four species showed substantial levels of antisense transcription covering 40 to 59% of protein-coding genes. By genome-wide mapping of TSS and PAS, our analysis revealed the similarities and differences of the regulatory sequences determining the initiation and polyadenylation of both mRNA and NATs among the four species, expanding our previous observations [32] from an evolutionary perspective.

Analyses of NATs in the *Entamoeba* core-genome revealed a set of genes ($n = 513$) that are associated with NATs in at least two species, and are involved in cell motility, nuclear structure, cell cycle, gene transcription and transport and nucleotide metabolism, coenzymes and amino acids metabolism. Among them, 86 genes are commonly associated with NATs in all four species, representing promising loci for further studies of NAT biogenesis. Differential expression analyses of these mRNA/NAT pairs in the core-genome revealed the interdependence between mRNA and NAT expression (Table S11), implying the potential regulatory functions of NAT or the potential linkage between mRNA and NAT biogenesis.

Interestingly, a substantial number of genes involved in lipid metabolism are associated with NAT up-regulation in *E. histolytica*. These encoded enzymes involved in

the degradation of lipids (lipases) and in the biogenesis of glycosylphosphatidylinositol (GPI), which is a moiety that anchors cell surface proteins to membranes, as well as kinases/phosphatases involved in the renewal of phospholipids. Multiple phospholipids play an essential role in vesicular membrane trafficking and pathogenicity of *E. histolytica* [67,68]. These results are in line with the vesicular traffic process; NAT up-regulation is also observed in genes encoding the structural components of the endoplasmic reticulum (ER) as well as small GTPase signalling factors, which are diverse, abundant in *E. histolytica* and regulate vesicular trafficking [69]. For example, small GTPases constitute signal platforms regulating endocytosis, ER protein and lipid translocation, phagocytosis and trogocytosis of human cells, as well as parasite motility.

In addition, the differential expression of NATs between *E. histolytica* and *E. moshkovskii* genes involving DNA damage and repair is particularly interesting as previous studies suggested that DNA recombination and allele reassortment should be more important than point mutations in *E. histolytica* genome [70]. In contrast, genomic divergence due to point mutations in *E. moshkovskii* strains is much greater than that demonstrated for *E. histolytica* strains [14]. One may be tempted to hypothesize that the differential expression of NATs might alter the expression of genes involving DNA damage and repair and eventually affect the frequency of DNA recombination in *Entamoeba* spp. This highly speculative consequence of NAT differential expression is expected to occur through regulating the expression of Mre11 (or other enzymes involved in DNA repair) rather than through a global mechanism dependent on antisense transcription, because the high abundance of NAT observed in this work does not correlate with the relatively low rate of recombination in *Entamoeba* species.

Although we demonstrated the existence of NAT and the interdependence of mRNA and NAT expression (Table S11) in the four species of *Entamoeba*, we are far from understanding their biogenesis and potential roles in gene regulation. In many cases, we do not observe positive correlation between NATs and mRNA expression (i.e., discordant pairs), and these discordant and concordant relationships between mRNA/NAT pairs suggest that there could be diverse mechanisms underlying their biogenesis or their potential regulatory roles in gene expression. Data from other eukaryotes indicate that NATs may regulate gene expression at the post-transcriptional or translational levels including modifying epigenetic markers [71] or specific silencing of genes [72]. In addition, antisense transcription also exists in other protozoan parasites. For example, in *Plasmodium*, NATs are associated with ~24% of coding genes, with a strong clustering towards the 3′ end of CDSs, and for some genes NAT transcript levels correlate with mRNA levels of neighbouring genes [73]. In *Toxoplasma gondii*, NATs are associated with ~21% of coding genes [74]. In *Giardia lamblia*, ~20% of sequenced cDNAs are NATs [75]. Although other examples exist [76], currently there are no clear functional roles for NATs in gene regulation in protozoan organisms. A promising recent finding shows that NAT species could represent a self-regulating mechanism of *T. gondii* gene expression at the post-transcriptional level [77]. In addition, in some fungi up to 50% of CDSs are associated with NATs [78] and it has been described that diverse phenomena such as transcriptional interference, chromatin remodelling and dsRNA formation are among the mechanisms involved in fungal gene regulation mediated by NATs [79]. Interestingly RNA interference pathways exist in *E. histolytica* and *E. invadens* [44,80]. Much remains to be learned about whether and to what extent antisense transcription in *Entamoeba* drives gene expression.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/article/10.3390/microorganisms10020396/s1, Figure S1: Small RNA linked to genes in the different species of *Entamoeba*, Figure S2: Boxplots illustrating the comparison between the specific 86 NAT associated genes and the other genes in each species (Wilcoxon-test results are bracketed), Figure S3: Genome mapping of NATs targeting syntenic genes common to the four species of *Entamoeba*, Figure S4: Mapping of NAT-targeted genes in three DNA fragments of *E. histolytica* contig DS571145, Table S1: Datasets, Table S2: Transcriptome assemblies, Table S3: Gene

lists, Table S4: sRNA genes and squared-chi tests, Table S5: TSS, Table S6: PAS, Table S7: EggNOG categories, Table S8: GO enrichment, Table S9: NAT and genome contigs, Table S10: Interspecies differential expression, Table S11: Differential expression squared-chi tests, Table S12A: GO enrichment *E. histolytica* vs. *E. dispar* up-regulated, Table S12B: GO enrichment *E. histolytica* vs. *E. dispar* down-regulated, Table S13A: GO enrichment *E. histolytica* vs. *E. moshkovskii* upregulated, Table S13B: GO enrichment *E. histolytica* vs. *E. moshkovskii* down regulated, Table S14A: GO enrichment *E. histolytica* vs. *E. invadens* upregulated, Table S14B: GO enrichment *E. histolytica* vs. *E. invadens* down regulated.

Author Contributions: C.W., M.K., J.-Y.C. and C.G.C. performed the wet experiments. D.M., C.-C.H. and N.G. designed the study. D.M., C.-C.H. and M.-A.D. performed the bioinformatics analysis. D.M., C.-C.H. and N.G. wrote the manuscript. J.-Y.C., M.-A.D. and C.G.C. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The work was funded by the National French Research Agency (ANR- GENOM-BTV 2010- Grant GENM-0011-01, GENAMIBE). The Transcriptome and Epigenome Platform is a member of the France Génomique consortium (ANR10-NBS-09-08).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available in the SRA database (https://www.ncbi.nlm.nih.gov/sra/) under accession number PRJNA781395.

Acknowledgments: The authors gratefully acknowledge the help of Hugo Varet (Hub de Bioinformatique et Biostatistique–DBC, Institut Pasteur) for his advice on differential expression experiments and to Susanne Warrenfeltz (AmoebaDB) for helping us in the search for synteny.

Conflicts of Interest: The authors have declared that no competing interests exist.

References

1. Stensvold, C.R.; Lebbad, M.; Victory, E.L.; Verweij, J.J.; Tannich, E.; Alfellani, M.; Legarraga, P.; Clark, C.G. Increased sampling reveals novel lineages of Entamoeba: Consequences of genetic diversity and host specificity for taxonomy and molecular detection. *Protist* **2011**, *162*, 525–541, doi:10.1016/j.protis.2010.11.002.
2. Stensvold, C.R. Pinning down the role of common luminal intestinal parasitic protists in human health and disease-status and challenges. *Parasitology* **2019**, *146*, 695–701, doi:10.1017/S0031182019000039.
3. Lokmer, A.; Aflalo, S.; Amougou, N.; Lafosse, S.; Froment, A.; Tabe, F.E.; Poyet, M.; Groussin, M.; Said-Mohamed, R.; Ségurel, L. Response of the human gut and saliva microbiome to urbanization in Cameroon. *Sci. Rep.* **2020**, *10*, 2856, doi:10.1038/s41598-020-59849-9.
4. Marie, C.; Petri, W.A. Regulation of virulence of Entamoeba histolytica. *Annu Rev. Microbiol* **2014**, *68*, 493–520, doi:10.1146/annurev-micro-091313-103550.
5. Cui, Z.; Li, J., Chen, Y.; Zhang, L. Molecular epidemiology, evolution, and phylogeny of Entamoeba spp. *Infect. Genet. Evol.* **2019**, *75*, 104018, doi:10.1016/j.meegid.2019.104018.
6. Clark, C.G.; Diamond, L.S. The Laredo strain and other 'Entamoeba histolytica-like' amoebae are Entamoeba moshkovskii. *Mol. Biochem. Parasitol.* **1991**, *46*, 11–18, doi:10.1016/0166-6851(91)90194-b.
7. Heredia, R.D.; Fonseca, J.A.; López; M.C. Entamoeba moshkovskii perspectives of a new agent to be considered in the diagnosis of amebiasis. *Acta Trop.* **2012**, *123*, 139–145, doi:10.1016/j.actatropica.2012.05.012.
8. Eichinger, D. Encystation in parasitic protozoa. *Curr. Opin. Microbiol.* **2001**, *4*, 421–426, doi:10.1016/s1369-5274(00)00229-0.
9. Chia, M.-Y.; Jeng, C.-R.; Hsiao, S.-H.; Lee, A.-H.; Chen, C.-Y.; Pang, V.F. Entamoeba invadens myositis in a common water monitor lizard (Varanus salvator). *Vet. Pathol* **2009**, *46*, 673–676, doi:10.1354/vp.08-VP-0224-P-CR.
10. Loftus, B.; Anderson, I.; Davies, R.; Alsmark, U.C.M.; Samuelson, J.; Amedeo, P.; Roncaglia, P.; Berriman, M.; Hirt, R.P.; Mann, B.J.; et al. The genome of the protist parasite Entamoeba histolytica. *Nature* **2005**, *433*, 865–868, doi:10.1038/nature03291.
11. Clark, C.; Alsmark, U.; Tazreiter, M.; Saito-Nakano, Y.; Ali, V.; Marion, S.; Weber, C.; Mukherjee, C.; Bruchhaus, I.; Tannich, E.; et al. Structure and content of the Entamoeba histolytica genome. *Adv. Parasitol.* **2007**, *65*, 51–190, doi:10.1016/S0065-308X(07)65002-7.
12. Lorenzi, H.A.; Puiu, D.; Miller, J.R.; Brinkac, L.M.; Amedeo, P.; Hall, N.; Caler, E.V. New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information.

PLoS Negl. Trop. Dis. **2010**, *4*, e716, doi:10.1371/journal.pntd.0000716.
13. Aurrecoechea, C.; Barreto, A.; Brestelli, J.; Brunk, B.P.; Caler, E.V.; Fischer, S.; Gajria, B.; Gao, X.; Gingle, A.; Grant, G.R.; et al. AmoebaDB and MicrosporidiaDB: Functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res.* **2011**, *39*, D612-619, doi:10.1093/nar/gkq1006.
14. Wilson, I.W.; Weedall, G.D.; Lorenzi, H.; Howcroft, T.; Hon, C.-C.; Deloger, M.; Guillén, N.; Paterson, S.; Clark, C.G.; Hall, N. Genetic Diversity and Gene Family Expansions in Members of the Genus Entamoeba. *Genome Biol. Evol.* **2019**, *11*, 688–705, doi:10.1093/gbe/evz009.
15. Wang, Z.; Samuelson, J.; Clark, C.G.; Eichinger, D.; Paul, J.; Van Dellen, K.; Hall, N.; Anderson, I.; Loftus, B. Gene discovery in the Entamoeba invadens genome. *Mol. Biochem. Parasitol.* **2003**, *129*, 23–31, doi:10.1016/s0166-6851(03)00073-2.
16. Ehrenkaufer, G.M.; Weedall, G.D.; Williams, D.; Lorenzi, H.A.; Caler, E.; Hall, N.; Singh, U. The genome and transcriptome of the enteric parasite Entamoeba invadens, a model for encystation. *Genome Biol.* **2013**, *14*, R77, doi:10.1186/gb-2013-14-7-r77.
17. Zaki, M.; Meelu, P.; Sun, W.; Clark, C.G. Simultaneous differentiation and typing of Entamoeba histolytica and Entamoeba dispar. *J. Clin. Microbiol* **2002**, *40*, 1271–1276, doi:10.1128/JCM.40.4.1271-1276.2002.
18. Willhoeft, U.; Buss, H.; Tannich, E. Genetic differences between Entamoeba histolytica and Entamoeba dispar. *Arch. Med. Res.* **2000**, *31*, S254, doi:10.1016/s0188-4409(00)00135-1.
19. Lorenzi, H.; Thiagarajan, M.; Haas, B.; Wortman, J.; Hall, N.; Caler, E. Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species. *BMC Genom.* **2008**, *9*, 595, doi:10.1186/1471-2164-9-595.
20. Kumari, V.; Sharma, R.; Yadav, V.P.; Gupta, A.K.; Bhattacharya, A.; Bhattacharya, S. Differential distribution of a SINE element in the Entamoeba histolytica and Entamoeba dispar genomes: Role of the LINE-encoded endonuclease. *BMC Genom.* **2011**, *12*, 267, doi:10.1186/1471-2164-12-267.
21. Pritham, E.J.; Feschotte, C.; Wessler, S.R. Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. *Mol. Biol. Evol.* **2005**, *22*, 1751–1763, doi:10.1093/molbev/msi169.
22. Clark, C.G.; Ali IK, M.; Zaki, M.; Loftus, B.J.; Hall, N. Unique organisation of tRNA genes in Entamoeba histolytica. *Mol. Biochem Parasitol* **2006**, *146*, 24–29, doi:10.1016/j.molbiopara.2005.10.013.
23. Tawari, B.; Ali, I.K.M.; Scott, C.; Quail, M.A.; Berriman, M.; Hall, N.; Clark, C.G. Patterns of evolution in the unique tRNA gene arrays of the genus Entamoeba. *Mol. Biol Evol* **2008**, *25*, 187–198, doi:10.1093/molbev/msm238.
24. Hackney, J.A.; Ehrenkaufer, G.M.; Singh, U. Identification of putative transcriptional regulatory networks in Entamoeba histolytica using Bayesian inference. *Nucleic Acids Res.* **2007**, *35*, 2141–2152, doi:10.1093/nar/gkm028.
25. Ehrenkaufer, G.M.; Haque, R.; Hackney, J.A.; Eichinger, D.J.; Singh, U. Identification of developmentally regulated genes in Entamoeba histolytica: Insights into mechanisms of stage conversion in a protozoan parasite. *Cell Microbiol* **2007**, *9*, 1426–1444, doi:10.1111/j.1462-5822.2006.00882.x.
26. De Cádiz, A.E.; Jeelani, G.; Nakada-Tsukui, K.; Caler, E.; Nozaki, T. Transcriptome analysis of encystation in Entamoeba invadens. *PLoS ONE* **2013**, *8*, e74840, doi:10.1371/journal.pone.0074840.
27. Manna, D.; Ehrenkaufer, G.M.; Singh, U. Regulation of gene expression in the protozoan parasite Entamoeba invadens: Identification of core promoter elements and promoters with stage-specific expression patterns. *Int. J. Parasitol.* **2014**, *44*, 837–845, doi:10.1016/j.ijpara.2014.06.008.
28. Naiyer, S.; Kaur, D.; Ahamad, J.; Singh, S.S.; Singh, Y.P.; Thakur, V.; Bhattacharya, A.; Bhattacharya, S. Transcriptomic analysis reveals novel downstream regulatory motifs and highly transcribed virulence factor genes of Entamoeba histolytica. *BMC Genom.* **2019**, *20*, 206, doi:10.1186/s12864-019-5570-z.
29. Zhang, H.; Ehrenkaufer, G.M.; Pompey, J.M.; Hackney, J.A.; Singh, U. Small RNAs with 5′-polyphosphate termini associate with a Piwi-related protein and regulate gene expression in the single-celled eukaryote Entamoeba histolytica. *PLoS Pathog.* **2008**, *4*, e1000219, doi:10.1371/journal.ppat.1000219.
30. Zhang, H.; Ehrenkaufer, G.M.; Manna, D.; Hall, N.; Singh, U. High Throughput Sequencing of Entamoeba 27nt Small RNA Population Reveals Role in Permanent Gene Silencing But No Effect on Regulating Gene Expression Changes during Stage Conversion, Oxidative, or Heat Shock Stress. *PLoS ONE* **2015**, *10*, e0134481, doi:10.1371/journal.pone.0134481.
31. Suresh, S.; Ehrenkaufer, G.; Zhang, H.; Singh, U. Development of RNA Interference Trigger-Mediated Gene Silencing in Entamoeba invadens. *Infect. Immun* **2016**, *84*, 964–975, doi:10.1128/IAI.01161-15.
32. Mornico, D.; Hon, C.-C.; Koutero, M.; Weber, C.; Coppee, J.-Y.; Dillies, M.-A.; Guillen, N. Genomic determinants for initiation and length of natural antisense transcripts in Entamoeba histolytica. *Sci. Rep.* **2020**, *10*, 20190, doi:10.1038/s41598-020-77010-4.
33. Zamorano, A.; López-Camarillo, C.; Orozco, E.; Weber, C.; Guillen, N.; Marchat, L.A. In silico analysis of EST and

genomic sequences allowed the prediction of cis-regulatory elements for Entamoeba histolytica mRNA polyadenylation. *Comput. Biol. Chem.* **2008**, *32*, 256–263, doi:10.1016/j.compbiolchem.2008.03.019.
34. Hon, C.-C.; Weber, C.; Sismeiro, O.; Proux, ,C.; Koutero, M.; Deloger, M.; Das, S.; Agrahari, M.; Dillies, M.-A.; Jagla, B.; et al. Quantification of stochastic noise of splicing and polyadenylation in Entamoeba histolytica. *Nucleic Acids Res.* **2013**, *41*, 1936–1952, doi:10.1093/nar/gks1271.
35. Mar-Aguilar, F.; Treviño, V.; Salinas-Hernández, J.E.; Taméz-Guerrero, M.M.; Barrón-González, M.P.; Morales-Rubio, E.; Treviño-Neávez, J.; Verduzco-Martínez, J.A.; Morales-Vallarta, M.R.; Resendez-Pérez, D. Identification and characterization of microRNAs from Entamoeba histolytica HM1-IMSS. *PLoS ONE* **2013**, *8*, e68202, doi:10.1371/journal.pone.0068202.
36. Saha, A.; Bhattacharya, S.; Bhattacharya, A. Serum stress responsive gene EhslncRNA of Entamoeba histolytica is a novel long noncoding RNA. *Sci. Rep.* **2016**, *6*, 27476, doi:10.1038/srep27476.
37. Clark, C.G.; Diamond, L.S. Methods for cultivation of luminal parasitic protists of clinical importance. *Clin. Microbiol Rev.* **2002**, *15*, 329–341, doi:10.1128/CMR.15.3.329-341.2002.
38. Criscuolo, A.; Brisse, S. AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **2013**, *102*, 500–506, doi:10.1016/j.ygeno.2013.07.011.
39. Dobin, A.; Gingeras, T.R. Mapping RNA-seq Reads with STAR. *Curr. Protoc. Bioinform.* **2015**, *51*, 11.14.1–11.14.19, doi:10.1002/0471250953.bi1114s51.
40. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079, doi:10.1093/bioinformatics/btp352.
41. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652, doi:10.1038/nbt.1883.
42. Wu, T.D.; Watanabe, C.K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **2005**, *21*, 1859–1875, doi:10.1093/bioinformatics/bti310.
43. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930, doi:10.1093/bioinformatics/btt656.
44. Zhang, H.; Ehrenkaufer, G.M.; Hall, N.; Singh, U. Identification of oligo-adenylated small RNAs in the parasite Entamoeba and a potential role for small RNA control. *BMC Genom.* **2020**, *21*, 879, doi:10.1186/s12864-020-07275-6.
45. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009, *10*, R25, doi:10.1186/gb-2009-10-3-r25.
46. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinform.* **2014**, *47*, 11.12.1–34, doi:10.1002/0471250953.bi1112s47.
47. Bullard, J.H.; Purdom, E.; Hansen, K.D.; Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform.* **2010**, *11*, 94, doi:10.1186/1471-2105-11-94.
48. Dugar, G.; Herbig, A.; Förstner, K.U.; Heidrich, N.; Reinhardt, R.; Nieselt, K.; Sharma, C.M. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet.* **2013**, *9*, e1003495, doi:10.1371/journal.pgen.1003495.
49. Bailey, T.L. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **2011**, *27*, 1653–1659, doi:10.1093/bioinformatics/btr261.
50. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 238, doi:10.1186/s13059-019-1832-y.
51. Wagner, G.P.; Kin, K.; Lynch, V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **2012**, *131*, 281–285, doi:10.1007/s12064-012-0162-3.
52. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550, doi:10.1186/s13059-014-0550-8.
53. Varet, H.; Brillet-Guéguen, L.; Coppée, J.-Y.; Dillies, M.-A. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS ONE* **2016**, *11*, e0157022, (2016), doi:10.1371/journal.pone.0157022.
54. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*; R Core Team: Vienna, Austria, 2019.
55. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2016.
56. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.K.; Cook, H.V.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology

resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309–D314, doi:10.1093/nar/gky1085.
57. Zhang, H.; Ehrenkaufer, G.M.; Hall, N.; Singh, U. Small RNA pyrosequencing in the protozoan parasite Entamoeba histolytica reveals strain-specific small RNAs that target virulence genes. *BMC Genom.* **2013**, *14*, 53, doi:10.1186/1471-2164-14-53.
58. North, J.A.; Wildenthal, J.A.; Erb, T.J.; Evans, B.S.; Byerly, K.M.; Gerlt, J.A.; Tabita, F.R. A bifunctional salvage pathway for two distinct S-adenosylmethionine by-products that is widespread in bacteria, including pathogenic Escherichia coli. *Mol. Microbiol* **2020**, *113*, 923–937, doi:10.1111/mmi.14459.
59. Nakada-Tsukui, K.; Saito-Nakano, Y.; Ali, V.; Nozaki, T. A retromerlike complex is a novel Rab7 effector that is involved in the transport of the virulence factor cysteine protease in the enteric protozoan parasite Entamoeba histolytica. *Mol. Biol Cell* **2005**, *16*, 5294–5303, doi:10.1091/mbc.e05-04-0283.
60. Tsuchiya, H.; Ohtake, F.; Arai, N.; Kaiho, A.; Yasuda, S.; Tanaka, K.; Saeki, Y. In Vivo Ubiquitin Linkage-type Analysis Reveals that the Cdc48-Rad23/Dsk2 Axis Contributes to K48-Linked Chain Specificity of the Proteasome. *Mol. Cell* **2017**, *66*, 488–502, doi:10.1016/j.molcel.2017.04.024.
61. Bodnar, N.; Kim, K.H.; Ji, Z.; Wales, T.E.; Svetlov, V.; Nudler, E.; Engen, J.R.; Walz, T.; Rapoport, T.A. Structure of the Cdc48 ATPase with its ubiquitin-binding cofactor Ufd1-Npl4. *Nat. Struct. Mol. Biol.* **2018**, *25*, 616–622, doi:10.1038/s41594-018-0085-x.
62. Bays, N.W.; Hampton, R.Y. Cdc48-Ufd1-Npl4: Stuck in the middle with Ub. *Curr Biol 2002,* **12**, R366-371, doi:10.1016/s0960-9822(02)00862-x.
63. Fu, M.; Blackshear, P.J. RNA-binding proteins in immune regulation: A focus on CCCH zinc finger proteins. *Nat. Rev. Immunol.* **2017**, *17*, 130–143, doi:10.1038/nri.2016.129.
64. Kastenmayer, J.P.; Green, P.J. Novel features of the XRN-family in Arabidopsis: Evidence that AtXRN4, one of several orthologs of nuclear Xrn2p/Rat1p, functions in the cytoplasm. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 13985–13990, doi:10.1073/pnas.97.25.13985.
65. Weber, C.; Koutero, M.; Dillies, M.-A.; Varet, H.; Lopez-Camarillo, C.; Coppée, J.Y.; Hon, C.-C.; Guillén, N. Extensive transcriptome analysis correlates the plasticity of Entamoeba histolytica pathogenesis to rapid phenotype changes depending on the environment. *Sci. Rep.* **2016**, *6*, 35852, doi:10.1038/srep35852.
66. Wight, M.; Werner, A. The functions of natural antisense transcripts. *Essays Biochem.* **2013**, *54*, 91–101, doi:10.1042/bse0540091.
67. Nakada-Tsukui, K.; Watanabe, N.; Maehama, T.; Nozaki, T. Phosphatidylinositol Kinases and Phosphatases in Entamoeba histolytica. *Front. Cell Infect. Microbiol.* **2019**, *9*, 150, doi:10.3389/fcimb.2019.00150.
68. Watanabe, N.; Nakada-Tsukui, K.; Maehama, T.; Nozaki, T. Dynamism of PI4-Phosphate during Interactions with Human Erythrocytes in Entamoeba histolytica. *Microorganisms* **2020**, *8*, E1050, doi:10.3390/microorganisms8071050.
69. Saito-Nakano, Y.; Wahyuni, R.; Nakada-Tsukui, K.; Tomii, K.; Nozaki, T. Rab7D small GTPase is involved in phago-, trogocytosis and cytoskeletal reorganization in the enteric protozoan Entamoeba histolytica. *Cell Microbiol.* **2021**, *23*, e13267, doi:10.1111/cmi.13267.
70. Weedall, G.D.; Clark, C.G.; Koldkjaer, P.; Kay, S.; Bruchhaus, I.; Tannich, E.; Paterson, S.; Hall, N. Genomic diversity of the human intestinal parasite Entamoeba histolytica. *Genome Biol.* **2012**, *13*, R38, doi:10.1186/gb-2012-13-5-r38.
71. Wanowska, E.; Kubiak, M.R.; Rosikiewicz, W.; Makałowska, I.; Szcześniak, M.W. Natural antisense transcripts in diseases: From modes of action to targeted therapies. *Wiley Interdiscip. Rev. RNA* **2018**, *9*, doi:10.1002/wrna.1461.
72. Zinad, H.S.; Natasya, I.; Werner, A. Natural Antisense Transcripts at the Interface between Host Genome and Mobile Genetic Elements. *Front. Microbiol.* **2017**, *8*, 2292, doi:10.3389/fmicb.2017.02292.
73. Siegel, T.N.; Hon, C.-C.; Zhang, Q.; Lopez-Rubio, J.-J.; Scheidig-Benatar, C.; Martins, R.M.; Sismeiro, O.; Coppée, J.-Y.; Scherf, A. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. *BMC Genom.* **2014**, *15*, 150, doi:10.1186/1471-2164-15-150.
74. Radke, J.R.; Behnke, M.S.; Mackey, A.J.; Radke, J.B.; Roos, D.S.; White, M.W. The transcriptome of Toxoplasma gondii. *BMC Biol.* **2005**, *3*, 26, doi:10.1186/1741-7007-3-26.
75. Elmendorf, H.G.; Singer, S.M.; Nash, T.E. The abundance of sterile transcripts in Giardia lamblia. *Nucleic Acids Res.* **2001**, *29*, 4674–4683, doi:10.1093/nar/29.22.4674.
76. Militello, K.T.; Refour, P.; Comeaux, C.A.; Duraisingh, M.T. Antisense RNA and RNAi in protozoan parasites: Working hard or hardly working? *Mol. Biochem. Parasitol.* **2008**, *157*, 117–126, doi:10.1016/j.molbiopara.2007.10.004.
77. Fahim, A.; Afrin, F.; Wen, G.; Ananvoranich, S. Characterization of natural antisense transcripts arisen from the locus encoding Toxoplasma gondii ubiquitin-like protease. *Mol. Biochem. Parasitol.* **2020**, *240*, 111334, doi:10.1016/j.molbiopara.2020.111334.

78. Donaldson, M.E.; Ostrowski, L.A.; Goulet, K.M.; Saville, B.J. Transcriptome analysis of smut fungi reveals widespread intergenic transcription and conserved antisense transcript expression. *BMC Genom.* **2017**, *18*, 340, doi:10.1186/s12864-017-3720-8.
79. Donaldson, M.E.; Saville, B.J. Natural antisense transcripts in fungi. *Mol. Microbiol.* **2012**, *85*, 405–417, doi:10.1111/j.1365-2958.2012.08125.x.
80. Zhang, H.; Alramini, H.; Tran, V.; Singh, U. Nucleus-localized antisense small RNAs with 5′-polyphosphate termini regulate long term transcriptional gene silencing in Entamoeba histolytica G3 strain. *J. Biol. Chem.* **2011**, *286*, 44467–44479, doi:10.1074/jbc.M111.278184.