



**HAL**  
open science

# Genomic determinants for initiation and length of natural antisense transcripts in *Entamoeba histolytica*

Damien Mornico, Chung-Chau Hon, Mikael Koutero, Christian Weber,  
Jean-Yves Coppee, Marie-Agnes Dillies, Nancy Guillen

## ► To cite this version:

Damien Mornico, Chung-Chau Hon, Mikael Koutero, Christian Weber, Jean-Yves Coppee, et al.. Genomic determinants for initiation and length of natural antisense transcripts in *Entamoeba histolytica*. Scientific Reports, 2020, 10 (1), pp.20190. 10.1038/s41598-020-77010-4 . pasteur-03228542

**HAL Id: pasteur-03228542**

**<https://pasteur.hal.science/pasteur-03228542>**

Submitted on 18 May 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



OPEN

## Genomic determinants for initiation and length of natural antisense transcripts in *Entamoeba histolytica*

Damien Mornico<sup>4,7</sup>✉, Chung-Chau Hon<sup>1,2,6,7</sup>, Mikael Koutero<sup>3</sup>, Christian Weber<sup>1,2</sup>, Jean-Yves Coppee<sup>3</sup>, Marie-Agnes Dillies<sup>4</sup> & Nancy Guillen<sup>1,2,5</sup>✉

Natural antisense transcripts (NAT) have been reported in prokaryotes and eukaryotes. While the functions of most reported NATs remain unknown, their potentials in regulating the transcription of their counterparts have been speculated. *Entamoeba histolytica*, which is a unicellular eukaryotic parasite, has a compact protein-coding genome with very short intronic and intergenic regions. The regulatory mechanisms of gene expression in this compact genome are under-described. In this study, by genome-wide mapping of RNA-Seq data in the genome of *E. histolytica*, we show that a substantial fraction of its protein-coding genes (28%) has significant transcription on their opposite strand (i.e. NAT). Intriguingly, we found the location of transcription start sites or polyadenylation sites of NAT are determined by the specific motifs encoded on the opposite strand of the gene coding sequences, thereby providing a compact regulatory system for gene transcription. Moreover, we demonstrated that NATs are globally up-regulated under various environmental conditions including temperature stress and pathogenicity. While NATs do not appear to be consequences of spurious transcription, they may play a role in regulating gene expression in *E. histolytica*, a hypothesis which needs to be tested.

*Entamoeba* is a genus of unicellular eukaryote *Amoebozoa* that separated from animals and fungi lineages after the evolutionary divergence of plants. *Entamoeba* is devoid of mitochondria and is found in a wide range of animal hosts, including humans that are natural hosts of at least eight well-known species of *Entamoeba*<sup>1</sup>. *Entamoeba histolytica* is the amoeba parasite responsible for human amoebiasis, one of the major neglected infectious disease affecting millions of people worldwide<sup>2</sup>. In recent years, genomics research has led to the major advances in the understanding of the parasite and the disease. The genome of *E. histolytica* consists of a large number of repetitive elements (~19%) and has high AT content (~76%)<sup>3</sup>, rendering its complete assembly problematic. The current assembly is sized 20 Mbp, consists of 1498 scaffolds with 8201 predicted genes, of which the vast majority (~76%) does not contain introns<sup>4</sup>. In addition to the scarcity of intron, amoeba genes have relatively short open reading frames (ORF) (averaged ~389 amino acids), as compared to other unicellular eukaryotes, e.g. *Dictyostelium discoideum* (another *Amoebozoa*, 518 amino acids) and *Plasmodium falciparum* (*Apicomplexa* parasite, 761 amino acids)<sup>5</sup>. Genes with short ORFs is a hallmark of compact genomes, e.g. the microsporidian *Encephalitozoon cuniculi* genome encodes 2000 genes in 3 Mbp<sup>5</sup>. These data highlight the compactness of *E. histolytica* genome with ~1 gene in 2 kb, comparing to the human genome with 1 gene in ~120 kb. Based on genomic and cDNA sequences, the fundamental characteristics of mRNA processing has been inferred. For example, the average introns length in *E. histolytica* is 75 nucleotides with well-preserved splicing sites 5'-GUU UGUU—UAG 3'<sup>4,6,7</sup> and the sequence motif involved in mRNA polyadenylation includes an AU-rich motif within the consensus sequence UA(A/U)UU<sup>8</sup>.

Analysis of *E. histolytica* transcriptome using RNA sequencing (RNA-Seq) suggested that 98% of the annotated coding regions is transcribed<sup>9</sup>. Despite the functional annotations of *E. histolytica* genome remains incomplete,

<sup>1</sup>Unité Biologie Cellulaire du Parasitisme, Institut Pasteur, Paris, France. <sup>2</sup>Institut National de La Santé Et de La Recherche Médicale, INSERM U786, Paris, France. <sup>3</sup>Plate-forme Transcriptome et Epigénome, Institut Pasteur, Paris, France. <sup>4</sup>Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, CNRS USR 3756, Institut Pasteur, Paris, France. <sup>5</sup>Centre National de La Recherche Scientifique, CNRS ERL9195, Paris, France. <sup>6</sup>Present address: Laboratory for Genome Information Analysis, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. <sup>7</sup>These authors contributed equally: Damien Mornico and Chung-Chau Hon. ✉email: damien.mornico@pasteur.fr; nguillen@pasteur.fr

several studies have used transcriptomics to understand changes in gene expression in various processes: (1) during amoebic differentiation into cysts<sup>10</sup>, (2) according to different degrees of parasite virulence<sup>11–13</sup> or (3) as a result of trophozoite growth under various environmental constraints<sup>14,15</sup>. Based on these transcriptome data, several bioinformatics and experimental approaches successfully brought insight into activities in the amoebic genome. For example, diverse gene promoters have been identified regulating cell growth<sup>16</sup>, cyst formation<sup>10</sup> or transcription during infection<sup>17</sup>. Recently, the correlative analysis of the genome and of the mRNA revealed several repetitive short nucleotide sequences of DNA which, according to their number of copies, modulate the expression of genes<sup>18</sup>. Splicing sites and polyadenylation sites on mRNA have been also characterized at the genome-wide level<sup>19</sup>.

The recent advances in high-throughput sequencing technologies revealed the pervasive transcription from eukaryotic genomes, producing a wide varieties of RNA transcripts which do not coding for proteins (i.e. non-coding RNAs). However, the functionality of non-coding transcripts is, to date, subject to intense debate, determining whether these RNAs are simple products of transcriptional noise or they contribute to the regulation of genetic expression<sup>20</sup>. Analyses of eukaryotic genomes have also identified a surprisingly high proportion of overlapping gene pairs, which often involves a long non-coding RNA (lncRNA) transcribed from the opposite strands of another gene, termed natural antisense transcripts (NAT)<sup>21</sup>. Despite the prevalence of NATs, its regulatory potential on gene expression has been demonstrated only in a few cases<sup>22</sup>. Very little is known about these molecules in *E. histolytica*, although short noncoding regulatory RNAs have been described, including small antisense RNAs<sup>23</sup> and miRNAs<sup>24</sup>. In addition, an lncRNA has been identified in *E. histolytica*<sup>25</sup>, which is a polyadenylated transcript of 2.6 kb, transcribed by RNA polymerase II. Oxygen and heat stress increase expression levels of EhslnRNA indicating that it acts as general stress regulator. Given the compactness of *E. histolytica* genome, the transcribed lncRNAs, if any, are likely to overlap with another gene on the opposite strand, and their regulatory potential is worth exploring. To better understand the mechanisms of gene regulation in such a compact genome, we mapped transcript fragments, the transcription start sites (TSS) and polyadenylation sites (PAS) genome-wide and identified NAT pairs. We then quantified changes in the expression levels of NAT pairs in trophozoites harvested under growing conditions, environmental changes and during infection, aiming to explore the regulatory potential of NATs. Our data revealed the pervasiveness of NAT in *E. histolytica*, suggest that NAT transcription is not likely to be the sole consequence of spurious transcription and show that genomic sequences for TSS and polyA sites are similar for sense and antisense transcription. These facts indicate a transcriptional dynamic hitherto unknown in *E. histolytica*.

## Results

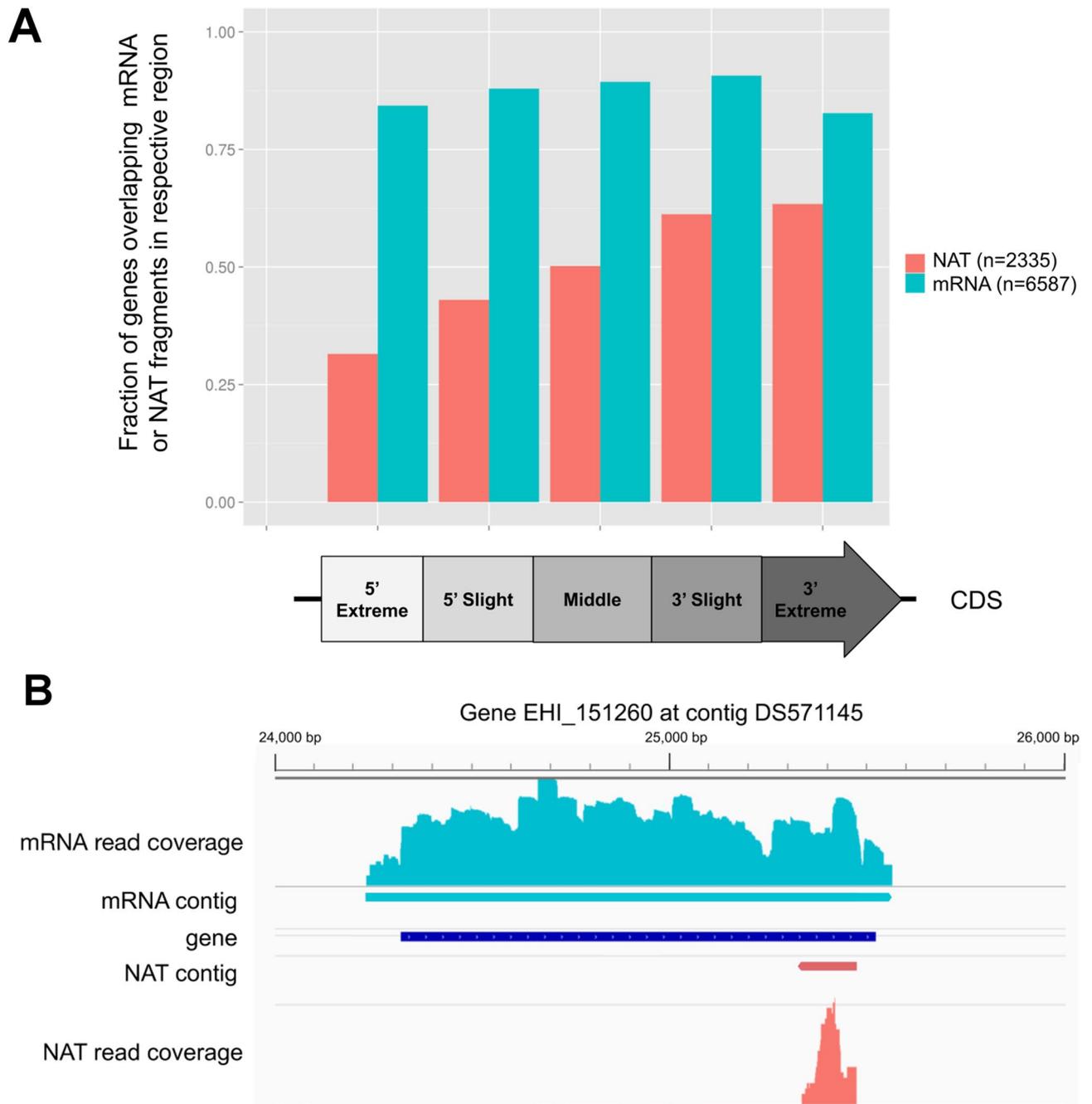
**Antisense transcription is abundant in *E. histolytica*.** To characterize the transcriptome of *E. histolytica*, we sequenced poly(A)+RNA of strain HM1:IMSS (n = 1 × 4, dataset 1A) (Table S1) and performed de novo assembly of transcriptome. We obtained 27,139 contigs (i.e. transcribed fragments) with a mean length of 327 nt (from 100 to 3881nt). About 63% of these transcribed fragments (n = 16,993, average 382nt) mapped to the same strand in 81% of genes (n = 6667), attesting a good coverage of the transcriptome. Another 14% of transcribed fragments (n = 3658, averaged 222nt) mapped to the opposite strand in 28% of the genes (n = 2335). We thus conclude at least a quarter of genes showed evidence of NAT transcription (Table S2). The validation of the detected NATs was performed through a northern blot for transcribed fragments that mapped to the opposite strand of one gene (EHI\_036570). We detected two NATs at this locus (after mapping the TSS and PAS, described below), which both were confirmed in the northern blot with expected molecular sizes (Fig. S1).

**NATs are biased towards the 3' end of their sense counterparts.** To understand the origin (i.e. biogenesis) of NATs in *E. histolytica*, we examined the overlap between annotated coding gene pairs on opposite strands. First, based on genome annotation, there were only 66 overlapping genes (Table S3). Second, only 2% of the NAT fragments (n = 73 of 3666) could be mapped to an adjacent gene on the same strand. This suggested that the majority of antisense transcription is not the consequence of read-through from adjacent genes. The contribution of overlapping coding gene pairs to the observed NATs in *E. histolytica* was rather negligible.

Several studies reported that antisense transcription could be correlated to sense gene splicing<sup>26</sup>. We investigated the association between antisense transcription and RNA splicing in *E. histolytica*. One-quarter of amoeba genes were predicted to be spliced (n = 2007) for a total number of 2559 of introns. By focusing on antisense transcription, we detected 151 introns (from 144 genes) overlapping, at least partially, with a NAT contig (example Fig. S2). Although splicing regulation by NAT could not be excluded, it did not seem to be a major control process. Indeed, approximately 7% intron harboring genes were associated with antisense transcription.

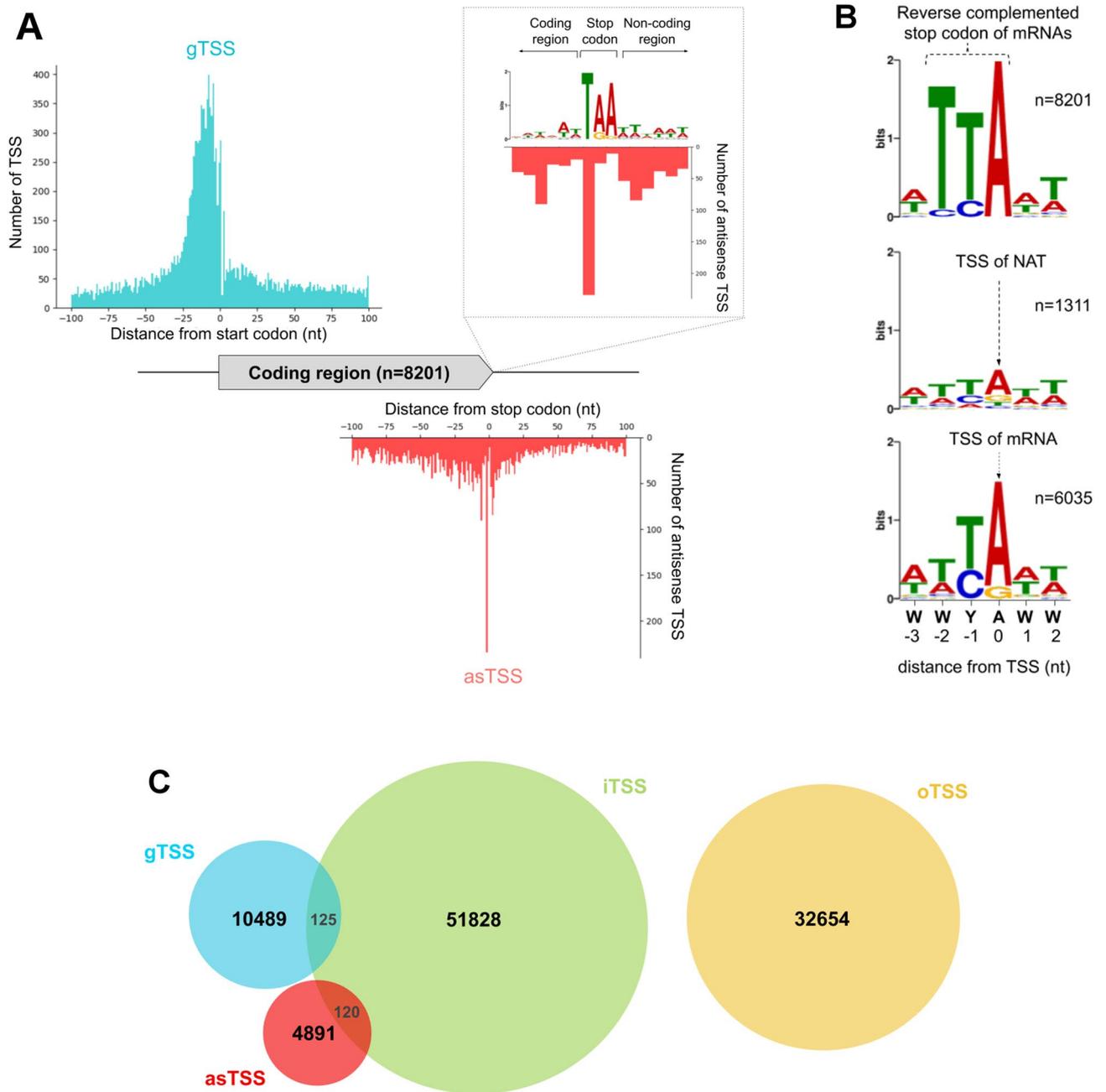
To shed light on the biogenesis of NATs, we investigated their pattern of coverage relative to their sense counterpart. We first divided the genomic regions of each coding gene into 5 equal sized bins and then counted the number of bins being covered by transcribed fragments on both strands. About 64% of the detected genes (n = 4257 of 6587) were fully covered (i.e. in all 5 bins) by transcribed fragments on the sense strand (i.e. mRNA), suggesting the de novo reconstructed transcriptome was reasonably well covered. We found only 12% of the NAT-possessing genes (n = 283 of 2335) were fully covered by transcribed fragments on the antisense strand (i.e. NAT). Most NATs were thus likely to cover partially their sense counterpart. In fact, we observed a bias of NAT coverage towards the 3' of its counterpart, with only ~30% of genes covered at 5' end comparing to ~60% of genes covered in 3' end (Fig. 1A). Based on these observations, we concluded that most of NATs are shorter than the mRNA and their transcription initiation is biased towards the 3' end of their sense counterpart (example in Fig. 1B).

**Transcription initiation of NATs at stop codon.** To get an insight into the initiation of NAT transcription, we mapped the TSS genome-wide using two libraries, one including a 5'-monophosphate dependent



**Figure 1.** NATs and CDS overlapping patterns. **(A)** Proportion of genes overlapping at least one mRNA (blue) or NAT (red) contig ( $n = 6667$  and  $n = 2335$  respectively) in the different regions of their CDS on both sides. 4257 genes were identified as fully covered by mRNA fragments and 283 by NAT. **(B)** Integrative Genome Viewer (IGV, <https://software.broadinstitute.org/software/igv/>)<sup>71</sup> screenshot for gene EHI\_151260 take as an example, representing RNA-seq coverage and assembly contigs for both sense (blue) and antisense (red) transcription.

terminator exonuclease (TEX) ( $n = 1$ , dataset 2A in Table S1) and a capped small RNA (csRNA) method ( $n = 1$ , dataset 2B in Table S1). Briefly, the TEX method<sup>27</sup> relies on the fact that 5' ends of primary transcripts which lack a 5'-monophosphate are protected from TEX degradation, while the csRNA method<sup>28</sup> is based on the fact that csRNAs (30 to 50nt) are produced at TSS through promoter-proximal pausing of Pol II<sup>29,30</sup>. While the mapping of a TSS of a transcript depends on its expression level in the TEX method, the level of Pol II pausing at the TSS (i.e. number of csRNA) is not strictly correlated to its expression level<sup>30</sup>. Therefore, these two methods are complementary. To control the level of background noise (i.e. false positive TSS derived from incomplete TEX digestion), we sequenced a control library with the same input RNA as in the TEX library but without TEX and TAP digestion, which should be unable to capture the TSS ( $n = 1$ , dataset 2C in Table S1). Background noise in both TEX and csRNA libraries were then subtracted, based on the ratio of reads in TEX or csRNA libraries to the control library (see “Methods”).



**Figure 2.** Transcription Start Sites identification for mRNA and NAT. **(A)** Mapping of TSS at CDS boundaries (with 100nt apart) on both strands, the region in the vicinity of the stop codon is magnified in the insert of the upper right corner. **(B)** Sequences logo computed around stop codon using the entire genome CDS reverse complemented (upper panel), asTSS (middle panel) and mRNA TSS (lower panel) found in this work. **(C)** Number of each type of TSS.

We then pooled the background subtracted signal (i.e. candidate TSS) from both TEX and csRNA dataset and clustered these candidate TSS (see “Methods”), yielding 95,326 TSS (i.e. TSS clusters, Table S4). These TSS were strongly enriched around the annotated start codons, peaking at 10nt upstream (Fig. 2A). This observation was consistent with the small size of intergenic regions (Fig. S3) and previous findings on the relative short 5’UTR in *Entamoeba* (~ 5 to 20 nt)<sup>31,32</sup>. Overall, the employed method has resulted in a successful and consistent capture of the TSS of mRNAs. In addition, about 73% of genes (n = 6035 of 8201) exhibited at least a TSS mapped within 100nt upstream of its start codon, suggesting a reasonable coverage of our TSS mapping. On the antisense strand, we observed a sharp peak of TSS surrounding the stop codon (Fig. 2A), consistent with the bias of NAT coverage towards gene 3’end as discussed earlier (Fig. 1).

Based on these observations, we classified all TSS into 4 classes (see “Methods”): 1) gTSS: “gene TSS” at proximity of start codon (100nt upstream) on the sense strand (n = 10,614 in 6035 genes), 2) asTSS: “antisense TSS” close to stop codons (+/- 100nt) on the antisense strand (n = 5011 in 2972 genes), 3) iTSS: “internal TSS” within

the ORF on same strand ( $n = 52,073$  in 6654 genes) and 4) oTSS: “orphan TSS” in between genes ( $n = 27,873$ ) (Fig. 2C). These oTSS may either correspond to unannotated ORFs located at contig boundaries of the draft genome<sup>33,34</sup>, or to possible artifacts. They were not further considered in the following analysis.

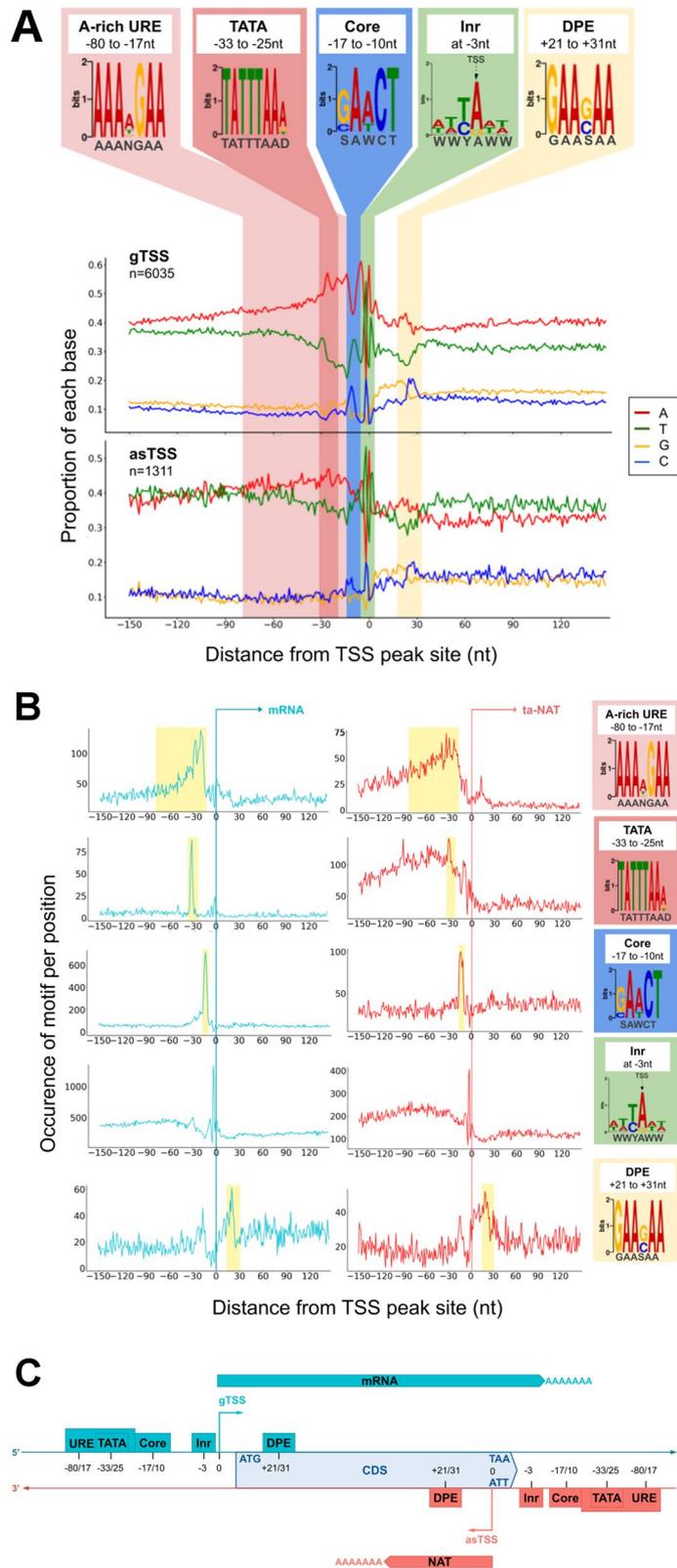
About 54% of the NAT-possessing genes ( $n = 1311$  of 2335) have at least one asTSS, implying at least a half of observed NAT was initiated within a 200nt window surrounding the stop codon (defined as terminal-associated NAT, ta-NAT). Thus, the stop codon represents a hotspot for antisense transcription initiation, and we indeed observed a sharp peak of TSS precisely located at the first base of the stop codon (Fig. 2A). To understand the genomic determinants for transcription initiation, we investigated the initiator motif (Inr) at the TSS of both mRNA and ta-NAT<sup>31</sup>. Analyses of TSS peak sequences revealed a conserved RNA Polymerase II (Pol II) Inr<sup>31</sup> for both mRNA and ta-NAT (i.e. WWYAWW, Fig. 2B), suggesting the majority of ta-NATs are transcribed by Pol II and the criteria for local selection of TSS by Pol II is similar in both mRNA and ta-NAT. To explain the observed preference of ta-NAT TSS at the stop codon (Fig. 2A), we compared the sequence around the stop codon (reverse complemented), with the Inr sequence of mRNA and ta-NAT (Fig. 2B). The reverse complement of the 6 nucleotides surrounding the stop codon (Fig. 2B upper panel), resembled the Inr motif (Fig. 2B middle and lower panel). In particular, the first base (i.e. T) and second base (i.e. T or C) of the stop codon was similar to the YA dinucleotide at -1 and 0 position of Inr motif. The ta-NAT TSS hotspot at stop codon might thus be a consequence of preferential TSS selection by Pol II, due to the Inr motif at stop codon on the antisense strand.

**NAT and mRNA share the same core promoter architecture.** We compared the sequence composition surrounding the TSS of mRNA (gTSS) and ta-NAT (asTSS) and found that they share similar characteristics (Fig. 3A), including (1) an A-rich region around -80 to -20 nt, (2) a C/T enriched region around -10 nt, (3) a YA motif around TSS (i.e. Inr in Fig. 2B), as well as (4) a C/G enriched region at +25nt. In addition, a T/A enriched region (at -30 nt) was found within the A-rich region in gTSS, and to less extent in asTSS. Motif enrichment analyses of gTSS sequences revealed the overrepresentation of 4 motifs (in addition to Inr in Fig. 2B), including an A-rich Upstream Regulatory Element (i.e. A-rich URE, AAANGAA,  $p = 5.5e-044$ ), a TATA-like box (i.e. TATA, TATTTAAD,  $p = 9.9e-050$ ), an upstream core motif (i.e. Core, SAWCT,  $p = 2.7e-588$ ) and a downstream promoter motif (i.e. DPE, GAASAA,  $p = 1.7e-019$ ) (Fig. 3A). It is noted that the TATA, Core and Inr described here are likely homologous to the previously described “non-consensus” TATA element, GAAC element, and initiator element respectively<sup>35</sup>. While the A-rich URE might be homologous to some of the “URES” described previously<sup>35</sup>, in this study, it refers to a loosely conserved A-rich motif broadly distributed between -80 to -20 nt, and the TATA-like box seems to be a specific variant of A-rich URE located sharply at -30 nt. The DPE described here is likely to be homologous to that of promoters in *Drosophila melanogaster*<sup>36</sup>. Analyses of motif occurrence showed that all of the 5 motifs were equally positionally enriched in both gTSS and asTSS (shaded ranges in Fig. 3B). These results clearly suggest that NAT and mRNA transcription takes place from core promoters with the same genomic architecture (Fig. 3C).

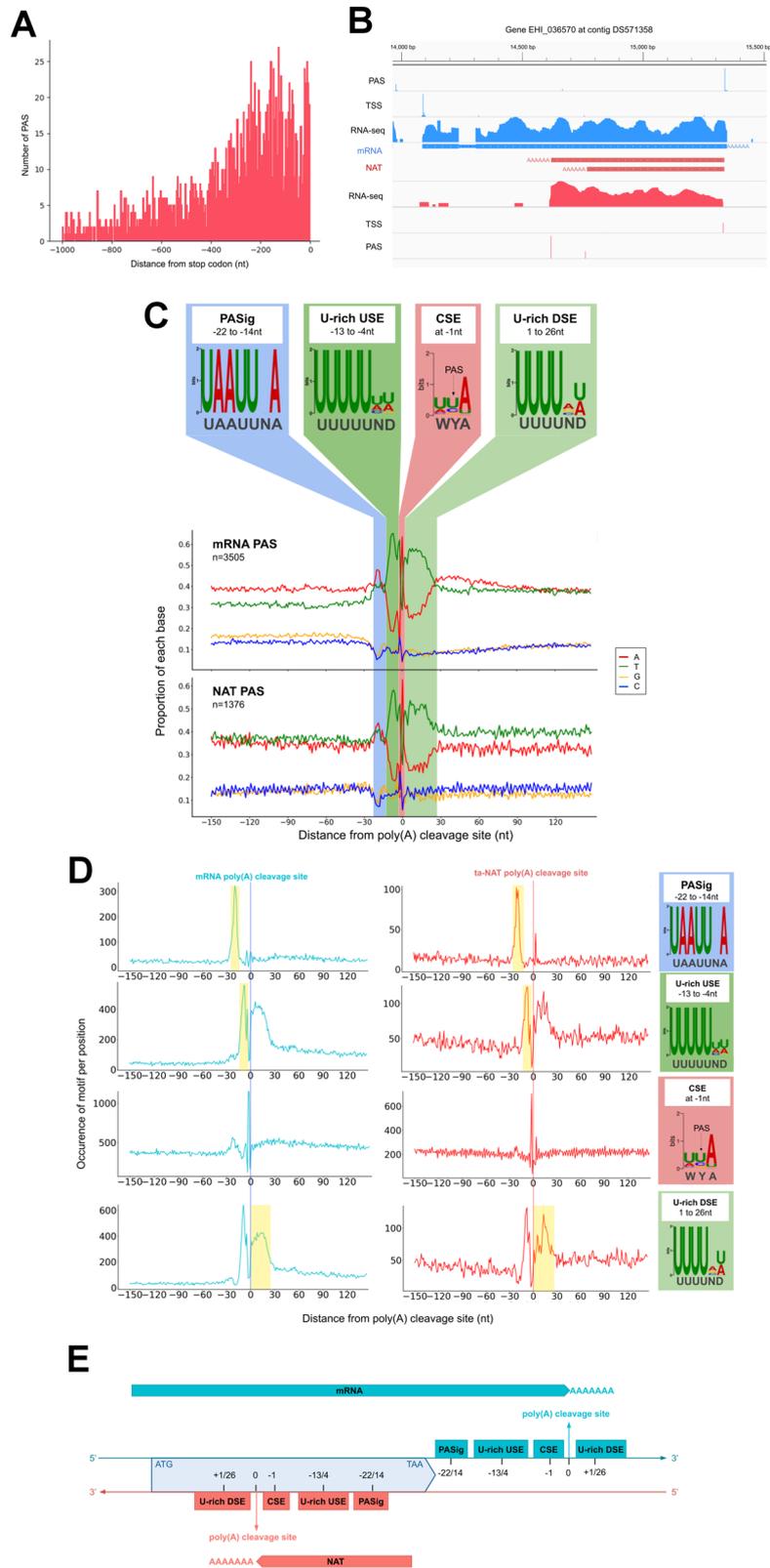
**Genomic determinants for polyadenylation and length of NATs.** As the above libraries were prepared from polyadenylated RNAs, we assumed the observed ta-NAT is polyadenylated. Next, we sought to investigate the genomic determinants of PAS of ta-NAT. We mapped the PAS genome-wide using the method we described previously<sup>19</sup>. Majority of identified PAS are located 20nt downstream of the stop codon, corresponding to the expected location of mRNA PAS<sup>19</sup>, suggesting most of the identified PAS are genuine. Closely located PAS were grouped into clusters as described previously<sup>19</sup>. Information of all PAS ( $n = 6518$ ) can be found in Table S5. Totally 2737 PAS were mapped to the antisense strand within the ORF of 1996 genes (Table S6). About 60% of these antisense PAS (883 of 1493, only genes with ORF > 1 kb were considered, Fig. 4A, Table S5) are located within 500 bp upstream of the stop codon, suggesting the majority of ta-NATs are polyadenylated within short distance after their transcription initiation from the 3' end of genes, generating relatively short ta-NAT transcripts. In some cases, we observed multiple antisense PAS within 1 kb upstream of stop codon ( $n = 523$  genes, Table S6), generating short ta-NATs of various lengths (an example in Fig. 4B, northern blot confirmation in Fig. S1). These data suggest, to some extent, ta-NATs could be multiple overlapping short transcripts initiated from the same TSS at gene 3' end, but ended with various lengths, depending on the choice of PAS.

Our previous analyses showed the conserved motifs surrounding PAS of eukaryotic mRNA are generally present in *E. histolytica* mRNA<sup>19</sup>. In fact, the sequence compositions surrounding the PAS of mRNA and ta-NAT were similar (Fig. 4C), suggesting these motifs are also present in PAS ta-NAT. Motif enrichment analysis suggested an overrepresentation of 3 motifs surrounding the PAS, (1) Polyadenylation Signal (PASig, UAAUUNA,  $p = 1.1e-238$ ), (2) U-rich Upstream Sequence Element (U-rich USE, UUUUUHD,  $p = 1.1e-574$ ) and (3) U-rich Downstream Sequence Element (U-rich DSE, UUUUNW,  $p = 6.2e-478$ ). We then defined a Cleavage Sequence Element (CSE, WYA) by aligning the 3 nucleotides around peak of mRNA PASC (Fig. 4C). The occurrence of these motifs surrounding the mRNA and ta-NAT PAS are very similar (Fig. 4D), suggesting the location of ta-NAT PAS depends on the same set of genomic determinants as those of mRNA, which are thus reverse encoded in the gene coding sequence (Fig. 4E).

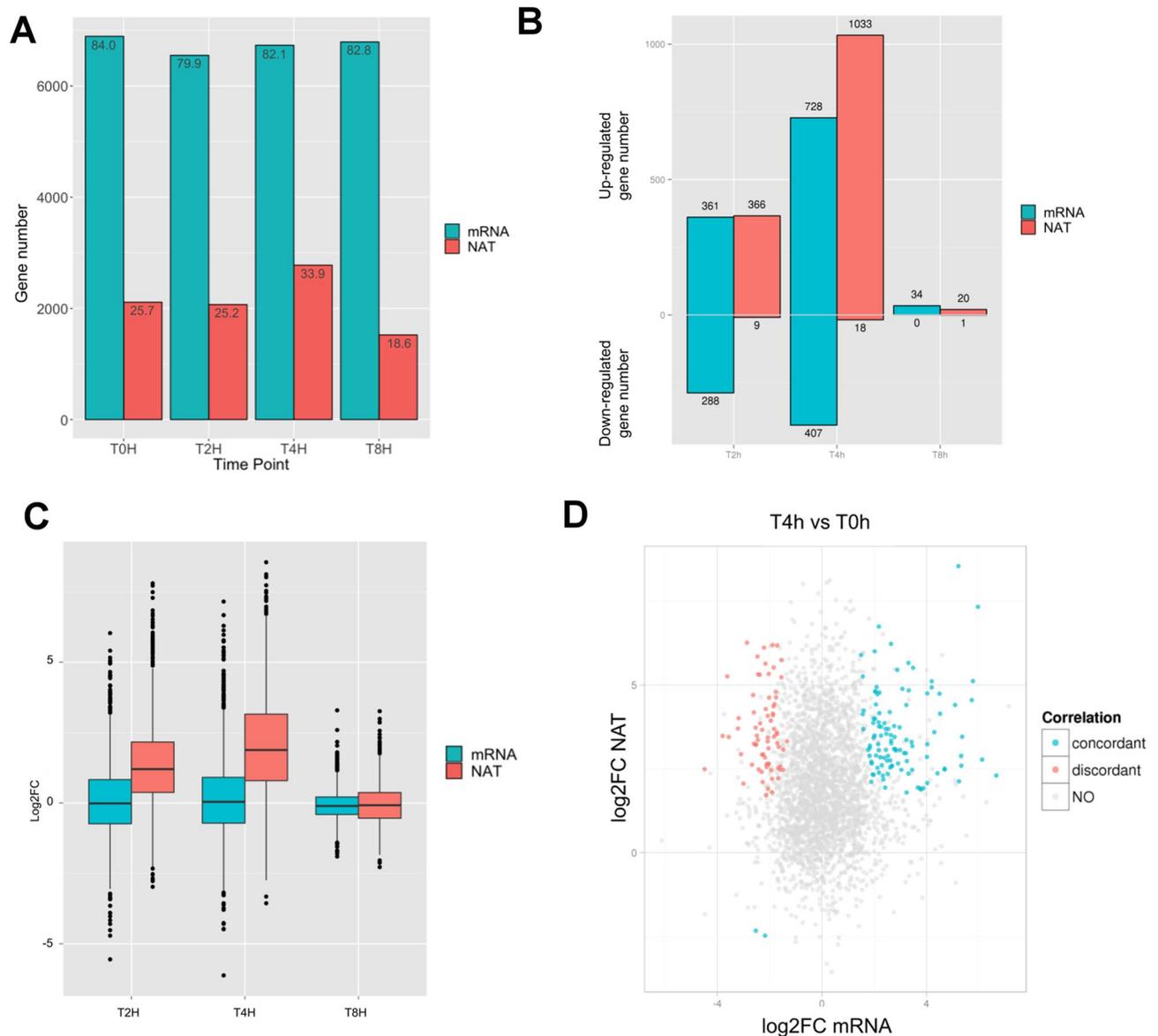
Based on these results, codon might be selected in order to favor the emergence of these motifs on the opposite strand. Considering the Relative Synonymous Codon Usage (RSCU), we found that codons with an A in the synonymous position is preferentially selected against the ones with a T for a same amino acid (Fig. S4). In line with this observation, the codon usage of *E. histolytica* might be evolutionarily selected to adapt to the enrichment of U-rich motifs in opposite strand and might represent an evolutionary strategy for compact genomes to attain extra regulatory components.



**Figure 3.** Transcription Start Sites motifs enrichment for mRNA and NAT. (A) Nucleotide sequence composition and motifs enrichment 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 (B) Appearance frequencies of each motif around TSS of both mRNA and NAT. (C) Schematic representation of TSS promoters positions on both DNA strands.



**Figure 4.** Polyadenylation sites features for mRNA and NAT. (A) Distribution of PAS distance from stop codon (position 0) on antisense DNA strand. (B) Integrative Genome Viewer (IGV, <https://software.broadinstitute.org/software/igv/>)<sup>71</sup> screenshot for gene EHI\_036570, representing RNA-seq coverage, TSS and PAS on both direct and reverse strand. (C) Nucleotide sequence composition and motif enrichment around the RNA cleavage site (position 0 corresponds to 1nt before the cleavage site) of both mRNA and NAT. (D) Appearance frequencies of each motif around the RNA cleavage site of both mRNA and NAT (position 0 corresponds to 1nt before the cleavage site). (E) Schematic representation of PAS patterns positions on both DNA strands.



**Figure 5.** Impact of temperature change on transcription. **(A)** Number of genes overlapped by mRNA and NAT contigs at different time-points of the experiment. **(B)** Proportion of up-regulated and down-regulated mRNA and NAT at T2h, T4h and T8h versus T0h, detailed in Table S8. **(C)** Log<sub>2</sub>FC box-plot for all genes at T2h, T4h and T8h versus T0h. **(D)** The fold change in gene expression (FC) was compared by plotting the log<sub>2</sub>FC of mRNA (x axis) versus log<sub>2</sub>FC of NAT (y axis) for each gene having at least one NAT contig identified, at T4h versus T0h. The color of points illustrates the differential expression type: none or unidirectional (grey), both concordant (blue), both discordant (red). Figure produced using R<sup>69</sup> and ggplot2<sup>70</sup>.

### Expression of NATs is regulated upon stress and between trophozoites of different virulence.

We assessed whether NAT expression was regulated by analyzing its transcription in different contexts. We profiled, using RNA-Seq, the transcriptomes of (1) trophozoites under stress (i.e. heat shock and recovery) and (2) trophozoites exhibiting different virulence. Here we first focus on their transcriptomes upon heat shock and recovery. Briefly, cells were passed from 37 to 42 °C, followed by a recovery stage at 37 °C. Samples were collected at 0 h (37 °C), 2 h (42 °C), 4 h (42 °C) and 8 h (37 °C) (dataset 3 in Table S1, “Methods”). We counted the number of genes with clear NAT transcription at each timepoint, by the presence of antisense transcribed fragments from de novo transcriptome assembly (Table S7, “Methods”). While the percentage of genes with mRNA transcribed fragments is relatively steady across time points (~79–82%), we observed a substantial increase of the percentage of genes with NAT transcribed fragments during heat shock at 4 h (~34%) followed by a decrease during recovery at 8 h (~19%) (Fig. 5A). The data suggested there was a general induction of NAT transcription under stress (during heat shock at 4 h), and the induced NAT transcription was reversible (during recovery at 8 h). This observation is consistent with differential expression analyses (“Methods”), for which we observed a sharp increased number of genes with significantly antisense transcription up-regulated (adjusted  $p$  value  $\leq 0.05$  with at least twofold-change, “methods”, Table S8) during heat shock at 4 hr (compared to 0 hr) (Fig. 5B,C).

Then, we explored the interaction between sense (i.e. mRNA) and antisense (i.e. NAT) transcription. We wondered whether a mRNA is more likely to be differentially expressed upon heat shock (i.e. dependent) when NAT is transcribed at the same locus. Focusing on 4 h, 1135 mRNAs were differentially expressed (728 up- and 407 down-regulated compared to 0 h) and 17% of them have NAT transcribed. A  $\chi^2$  test revealed a significant, although moderate, dependence of mRNA differential expression and NAT presence ( $\chi^2 = 5.4994$ ,  $df = 1$ ,  $p = 0.0190$ , Table S9—sheet 1), implying a potentially general impact of NAT in regulation of mRNA expression. However, we did not observe a global correlation between the fold-changes of mRNA and NAT ( $R^2 = 0.0018$ , Fig. 5D). Notably, there were 185 pairs of mRNA and NAT differentially expressed at the same time. 113 pairs changed in the same direction (i.e. concordant, *blue* in Fig. 5D) and 72 pairs changed in the opposite direction (i.e. discordant, *red* in Fig. 5D). Previous studies suggested the concordant (i.e. correlated) and discordant (i.e. anticorrelated) expression of mRNA and NAT pairs might have reflected different regulatory relationships with distinct molecular mechanisms, e.g. co-degradation in concordant expression and transcriptional interference in discordant expression<sup>37</sup>. While our data showed no systematically concordant or discordant expression between the mRNA and NAT pairs in *E. histolytica*, some genes display these peculiar patterns of transcription (Table S9—sheet 2). As examples, small GTPases; guanine nucleotide exchange factor GEF, cysteine proteases, chitinase and glycoprotein Jacob involved in encystation are among the concordant NAT- mRNA upregulated gene expression, whereas several GTPase activating protein GAP, two cysteine protease binding proteins and several enzymes are discordant.

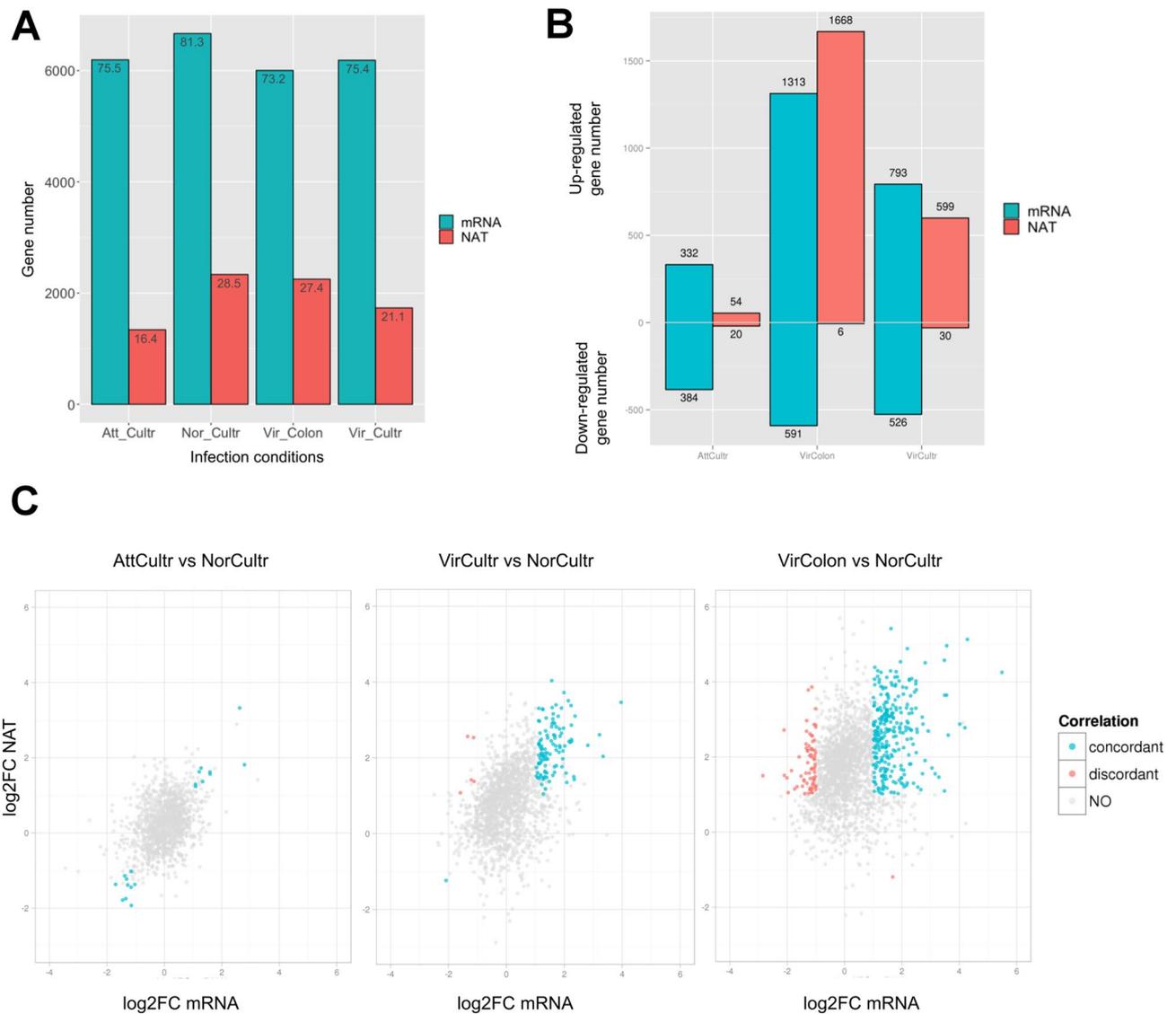
To further extend the above observations, we performed similar analyses on the transcriptomes of trophozoites of different virulence and in different infection conditions, as described previously<sup>12</sup> (dataset 1A–D in Table S1, “Methods”). Briefly, we profiled samples: (1) the “normal” trophozoites maintained in short-time culture (NorCultr), (2) virulent trophozoites freshly extracted from hamster with liver abscess (VirCultr), (3) virulent trophozoites cultured in human colon explants (VirColon), (4) attenuated trophozoites with long-time culture (AttCultr). First, we observed slight changes of NAT transcription across the 4 infection conditions (in terms of presence of antisense transcribed fragments as described) (Fig. 6A, Table S10). We next identified the genes differentially expressed in VirCultr, VirColon and AttCultr, compared to NorCultr (Fig. 6B, Table S11). The independence of both sense and antisense transcription, was tested using a  $\chi^2$  test based on the presence or the absence of associated NAT with differentially expressed or not modulated genes (Table S11). To refine this relation, we performed separately two analyses on differentially up and down expressed genes to figure out if the dependency was more significant in up or down regulation. We highlighted a dependence between presence of NAT and the up regulation, specifically in “Vir” and “VirColon” conditions ( $p = 7.925e-05$  and  $p = 1.48e-02$  respectively). Even if we could not exclude a dependency between NAT presence and up-regulation of mRNA in some genes due to a general increase of transcription on both strands, some regulatory mechanisms have already been demonstrated for common upregulation of transcripts sense and antisense<sup>38</sup>. Next, similar to the observations in heat shock, we did not observe a global trend of changes in mRNA and NAT co-expression, as illustrated in Fig. 6C. The majority of identified genes present a concordant pattern (Table S11, sheet 2), whereas was noticeable 71 genes in the condition VirColon versus NorCultr presenting a discordant pattern (NAT up-regulated and mRNA down-regulated); among them several enzymes can be highlighted (Table S11, sheet 3).

Finally, we identified a set of genes ( $n = 457$ ) exhibiting a significant NAT transcription in all 8 samples described above (Table S13, sheet 1). Among them 257 gene products were identified at UniProt library collection and corresponds to proteins grouped in 6 major categories including: cytoskeleton related; traffic, signaling and stress; enzymes; nucleic acid interactors and pathogenesis (Table S13, Sheet 1). There we highlighted proteins having a role in pathogenesis such as the Gal/GalNAc lectin intermediate subunits Igl1 and Igl2 (EHI\_006980, EHI\_065330) and the two light chain subunits EHI\_148790 and EHI\_058330. The amoebapore B, two proteinases, the serine rich protein, the 20 kDa factor and 10 members of the BspA family were also identified. Moreover, 93 identified genes are full length covered by NAT in trophozoites under normal culture conditions (Table S13, sheet 2).

## Discussion

Antisense transcription is a general feature of living entities and cells since it has been reported in viruses<sup>39</sup>, bacteria<sup>40</sup>, protozoa<sup>41</sup>, fungi<sup>42</sup>, plants<sup>43</sup>, invertebrates<sup>44</sup> and mammals<sup>45</sup>. Although the proportion of the observed NATs being functional remains elusive, a number of NATs are reported to modulate the expression of their cognate sense RNA through various mechanisms (for reviews see<sup>22,46</sup>). For example, such a modulation may arise epigenetically, through recruitment of Polycomb Repressive Complex 2 for gene-silencing by histone modification, or transcriptionally, through formation of RNA: RNA duplexes that affects alternative splicing, nuclear retention, RNA editing and transcript stability. Here we demonstrated that more than a quarter of the genes in *E. histolytica* showed significant evidence of NAT expression, which is far from negligible compared to other species<sup>47</sup> and equivalent to the proportion of NATs found in humans<sup>48</sup>. For instance, few NATs have been identified in *Dictyostelium discoideum*<sup>49</sup> and *Trypanosoma brucei*<sup>50</sup> (162 and 182 pairs respectively), while *Plasmodium falciparum* shows a similar ratio of antisense transcription (24% of protein-coding genes)<sup>41</sup>.

Based on genome-wide TSS mapping, we demonstrated the majority of *E. histolytica* NATs are transcribed around the 3' end of the ORFs, and in particular precisely initiated at the stop codon with recognizable promoter architecture. In addition, by genome-wide PAS mapping, we showed these NATs are polyadenylated and their PAS are reversely encoded in the corresponding ORF which has a deterministic effect on the length of NATs. These results suggest the majority of the observed NATs are not likely to be due to sequencing artefacts. More importantly, the encoding of NAT regulatory sequences (i.e. motifs associated with their TSS, PAS and Inr) on the opposite direction of the encoding DNA strand, unambiguously indicates for the first time the highly efficient usage of intergenic spaces in such a compact genome.



**Figure 6.** Infection condition effects on transcription. **(A)** Number of genes overlapped by mRNA and NAT contigs in different infection conditions of the experiment. **(B)** Proportion of up-regulated and down-regulated mRNA and NAT in AttCultr, VirColon and VirCultr versus NorCultr, detailed in Table S12. **(C)** The fold change in gene expression (FC) was compared by plotting the  $\log_2\text{FC}$  of mRNA (x axis) versus  $\log_2\text{FC}$  of NAT (y axis) for each gene having at least one NAT contig identified, in AttCultr, VirCultr or VirColon versus NorCultr. The color of points illustrates the differential expression type: none or unidirectional (grey), both concordant (blue), both discordant (red). Figure produced using R<sup>69</sup> and ggplot2<sup>70</sup>.

Interestingly, we observed a global up-regulation of NAT transcription during heat shock back to normal levels during recovery, suggesting the expression of NAT is physiologically regulated, despite their potential roles in regulation of gene expression remains elusive. The significant over-representation ( $\chi^2$  test,  $p < 0.05$ ) of NAT in genes with differentially expressed mRNAs implies some potential mechanisms of co-regulation, or co-biogenesis, between NAT and mRNA pairs. Intriguingly, we did not observe a global correlation, or anti-correlation, between the expression of NATs and their cognate mRNA. This lack of global correlation is similar to the NATs observed in *P. falciparum*<sup>41</sup>. Nonetheless, we cannot exclude the presence of distinctive and heterogeneous mechanisms<sup>38,47,51–53</sup> underlying the above observations, as we noticed some genes exhibiting clear concordant or discordant transcriptions. This study identified a set of 457 genes with confident NAT transcription across multiple experiments, representing a great opportunity for further experimental and in silico approaches to interrogate their roles in gene regulation, particularly for the ones associated with *E. histolytica* virulence, such as the Gal/GalNAc lectin subunits.

A strong correlation between the percentage of A/T bases of a genome and the number of identified anti-sense RNAs was demonstrated<sup>54</sup>. The high frequency of A/T increases the random occurrence of core promoter motifs, which are enriched with A/T. One could therefore speculate that NATs identified in *E. histolytica* (whose genome bear 76% of A/T bases) are largely derived from aleatory promoters motifs, and antisense transcription

might largely represent noise. However, we argue that their physiologically regulated expression under heat shock disadvantage this hypothesis. Also, while a wide distribution of TSS is expected if the emergence of NAT is random, instead, we observed a sharp TSS peak on stop codon. All the more so as this particular region has already been identified as a potential target for antisense regulation mechanism<sup>55</sup>. Overall, we concluded that paired with the compact genome constraints above discussed, these NATs offer great embedded opportunities for transcription regulation to *E. histolytica*.

## Methods

**Culture of *Entamoeba histolytica*.** *Entamoeba histolytica* strain HM1:IMSS was cultured in TYI-S-33 medium at 37 °C<sup>12</sup>.

**RNA isolation and northern blot analysis.** Trophozoites growing in TYI-S-33 medium were harvested and total RNA extracted using TRIZOL reagent (Invitrogen) and cleaned using the RNEasy cleanup kit (Qia-gen). Northern blots were performed according to the following protocols. Briefly, 10 µg or 20 µg of total parasite RNA (per electrophoresis lane) was denatured at 65 °C, loaded and molecules separated in a 1% agarose gel electrophoresis for 2 h at 95 V. RNA size standards were used to calibrate the samples (Ambion ref. 7170). RNA was transferred during 2 h to membrane filters (Ambion Bright star, ref. AM10102), the filters were crosslinked to the membrane by exposition to 1200 uJ and hybridized with biotinylated probes.

To prepare the RNA probes, DNA fragments of different lengths were PCR amplified from the amoebic genome according to EHI\_036570 gene sequence, using diverse primers according to the gene sequence and adding a T7 primer for in vitro amplification of the RNA (Fig. S1). PCR conditions using 1 µg of genomic DNA were: 94 °C for 3 min followed by 32 cycles of 94 °C for 30 s, 54 °C for 30 s and 72 °C for 45 s and then the PCR product was incubated for 5 min at 72 °C. In vitro transcription using biotinylated nucleotides and PCR fragments was done in the presence of T7 polymerase and Bio-11-UTP (Ambion ref. AM845) according to furnisher protocols (Ambion ref. AM 2082). The RNA product was treated by Turbo DNase, precipitated by ethanol and resuspended in distilled water, quantified with a NanoDrop and used for northern blotting experiments above described. RNAs were detected using the BrightStar BioDetect Kit (Ambion ref. AM1930) with prehybridization at 65 °C for 30 min, hybridization overnight at 62 °C and chemiluminescent alkaline phosphatase substrate revelation with the CDP-Star Substrate (ThermoFisher ref. T2145). Blots were exposed to film, subjected to autoradiography, scanned, and prepared for publication using Adobe Photoshop (version 7, San Jose, CA).

**Search for gene pairs in the genome of *E. histolytica*.** To examine the eventual presence of gene pairs on the genome of *E. histolytica* HM1: IMSS, we used the colocation tool from AmoebaDB (<https://amoebadb.org/amoeba/>). The steps were as follows: search for all genes in Taxonomy/organism (get an answer), click on Add a step, search for Taxonomy/organism and search and choose 1 relative to 2 in the overlapping diagram to combine the search results, click on continue. In the open window organize the colocation tool: click "genes from step 1" (small window in the left) and see the text whose exact region overlaps a gene from step 2 and is on "the opposite strand" (small window in the right). Click on submit. The tool proposes the overlapping genes on a new window.

**RNA extraction, library construction and sequencing.** Total RNA was extracted from approximately  $1 \times 10^6$  *E. histolytica* trophozoites (with each sample performed in triplicates) using Trizol (Invitrogen). The polyA fraction was purified from 10 to 100 µg of total RNA using Dynabeads according to the manufacturer's instructions (ThermoFisher ref. 28152103011150). Libraries were constructed using ScriptSeq mRNA-Seq Library Preparation Kit (Illumina) following manufacturer's recommendations and were quality controlled using Agilent Bioanalyzer. Sequencing was performed on a HiSeq 2000 (Illumina) to obtain 58 base single-end reads. All reads from different experiments were cleaned from adapter sequences with AlienTrimmer<sup>56</sup> (version 0.4.0) and low quality (<20) or short sequences (<20 nt) were removed from datasets.

**Genome reference and data repositories.** The reference genome was downloaded from AmoebaDB v34<sup>57</sup> (<https://amoebadb.org/common/downloads/release-44/EhistolyticaHM1IMSS/>). The total genome size is 20,80 Mbases, with a GC content of 24.2%, assembled in 1496 contigs, and 8201 annotated coding sequences. The data from experiments of TSS identification and heat-treated amoebae (dataset 2 and 3 in this paper) are available in the SRA database (<https://www.ncbi.nlm.nih.gov/sra/>) under the accession number PRJNA615171. The data concerning virulence conditions (dataset 1) derived from (Weber et al.)<sup>12</sup>, these data were already in the ArrayExpress database ([www.ebi.ac.uk/arrayexpress](http://www.ebi.ac.uk/arrayexpress)) under accession number E-MTAB-4882.

**Transcript contig construction.** Illumina libraries were merged within each dataset (Table S1) for NAT fragment analysis in order to reach bigger set of reads. Sequences were mapped to the reference genome of *E. histolytica*, using STAR v2.5.0<sup>58</sup> with a maximum intron length of 900 (–alignIntronMax), 2 mismatches maximum (–outFilterMismatchNmax), 10 locations maximum for a read mapping (–outFilterMultimapNmax), minimum overhang (–alignSJoverhangMin) of 25 for spliced alignments. Mapped reads were then split into direct and reverse strand sets with SAMTools<sup>59</sup>. Trinity assemblies<sup>60</sup> were performed on each set (direct and reverse) to construct transcript fragments, with a kmer size of 15nt (–KMER\_SIZE 15), a maximum intron size of 900nt (–genome\_guided\_max\_intron 900), a minimum coverage of 20 (–genome\_guided\_min\_coverage 20) and a minimum length of 100nt for each assembled contigs (–min\_contig\_length 100). Resulting transcript contigs were mapped on the reference genome using Gmap<sup>61</sup> with a maximum intron length of 900nt (–K 900). Lastly,

NAT fragment counting for each gene was produced by featureCounts<sup>62</sup> on the opposite strand (-s 2), with a minimum fraction of 10% of the contig length overlapping the gene, required for its assignment (-fracOverlap 0.1). Genes were counted as “NAT gene”, when at least 1 NAT contig was identified with featureCounts. In order to specify parts of genes preferentially covered by NAT contigs, we split the genes in 5 equal regions and counting was performed on each region as well, allowing contig assignment to all of their overlapping parts (-O -f).

**TSS identification and annotation.** Total RNA (10–100 µg) was polyA enriched twice using Sera-Mag Magnetic Particles Oligo(dT) coated (Thermo Scientific ref. 28152103011150) and fragmented 5 min (Ambion ref. AM8740). According to manufacturer’s instructions, Transcription Start Site enrichment was achieved with Terminator 5'-Phosphate-Dependant Exonuclease (TEBU ref. TER51020) followed by Tobacco Acid Pyrophosphatase (TEBU ref. T81050) treatment. For CapSeq, total RNA was similarly treated with an additional 15% TBE-urea gel size selection. RNA was treated with Terminator 5'-phosphate-dependent Exonuclease (Epicenter ref. TER51020), calf intestinal phosphatase (New England Biolab ref. M0290) and Tobacco Acid Pyrophosphatase (Epicenter ref. T19050). Libraries were constructed following manufacturer’s recommendations using TruSeq Small RNA Sample Prep Kit (Illumina, ref. RS-200-0012). The libraries were purified with AMPure XP beads (Agencourt ref. A63880) and controlled by Bioanalyzer DNA High Sensitivity Chips (Agilent ref. 5065-4626). Sequencing has been performed on a HiSeq 2000 (Illumina) in a multiplexed 51 + 7 bases single read using a TruSeq SR Cluster kit v2 cBot HS (Illumina, ref. GD-401–2510) and a TruSeq SBS kit v2 HS 50 cycles (Illumina, ref. FC-401-1002). After sequencing, files were generated using CASAVA 1.8.2 (Illumina). The TSS libraries yielded a total of 216,650,724 sequences reads for the csRNA library, 285,123,189 for the TEX library and 303,643,671 for the control library.

The reads were aligned to the reference genome of *E. histolytica*, using Bowtie version 0.12.7<sup>63</sup> with the following parameters: maximum 2 mismatches were allowed (-n 2) and reads mapped to multiple locations (-m 50) were reported only once (-k 1). The produced alignments were sorted and indexed with SAMTools<sup>59</sup>. Coverage graphs representing the numbers of mapped reads per nucleotide were generated based on the sorted reads using BEDTools<sup>64</sup>, focusing on 5' end position (-5). On each coverage an upper quartile normalization<sup>65</sup> was performed.

For each library, potential TSS were identified at the positions where all the following conditions were met ( $e_{Li}$  is the coverage at position  $i$  in the graph  $L$ ):

- Minimum coverage:  $e_{TEX(+)|csRNA}^i \geq 5$
- Minimum ratio:  $\frac{e_{TEX(+)|csRNA}^i}{e_{Control}^i} \geq 1.5$

TSS of both libraries were then merged, as the 2 methods were identified as consistent. TSS candidates within 10nts 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 cluster was then classified as “gene TSS” (gTSS), an “internal TSS” (iTSS), an “antisense TSS” (asTSS), or an “orphan TSS” (oTSS) if it could not be assigned to any of the previous classes<sup>33</sup>. A TSS cluster was classified as gTSS if it was located  $\leq 100$  bp upstream of a gene and as asTSS if it was located within the 200 bp surrounding the stop codons on the antisense strand. The TSS cluster with the strongest expression values (maximum peak height) among gTSSs of a gene was classified as primary (pTSS). “Internal TSS” (iTSS) were located within an annotated gene on the sense strand. TSS belonging to none of these conditions (intergenic TSS) were annotated as “orphan TSS” (oTSS).

**PAS identification and annotation.** First, reads with a stretch of five or more ‘A’ at their ends (or ‘T’ at their beginning) were selected for this analysis, as they potentially contain mRNA poly(A) tails. Redundant reads were removed and stretches of A at the ends were trimmed. Remaining reads with a minimum length of 18nt were then mapped on the reference genome using Bowtie<sup>63</sup> 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 (5nt up 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 with the following criteria: (1) reads with  $\geq 8$  nt within 10 nt immediate upstream of the PAS are A’s, (2) mapping with  $> 5$  nt immediate downstream of the PAS are A’s.

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 less than 2 reads of coverage at the peak were dismissed.

**Motif enrichment.** The sequences immediate upstream and downstream of the gTSS, and asTSS (100nt on each side), as well as the PAS of mRNA and NAT (50nt on each side) were used to scan for conserved motifs using DREME<sup>66</sup>. 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.

**Differential expression analysis.** Reads sequences of each replicate were mapped to the reference genome of *E. histolytica*, using STAR v2.5.0<sup>58</sup> with a maximum intron length of 900 (-alignIntronMax), 2 mismatches

maximum (`-outFilterMismatchNmax`), 10 locations maximum for a read mapping (`-outFilterMultimapNmax`), minimum overhang of 25 for spliced alignments.

Reads counting for each gene was produced by `featureCounts`<sup>62</sup> separately on the direct (`-s 1`) and opposite strand (`-s 2`), allowing multi-mapping reads (`-M`). Normalization was first calculated on the sense counting, and size-factors were then applied to both sense and antisense counting. Transcript differential expressions were calculated on the merged normalized counting (sense and antisense) using DESeq2 v1.24.0<sup>67</sup> within the SARTools pipeline v1.7.2<sup>68</sup>.

Analysis was conducted in R<sup>69</sup> and figures were produced using the package `ggplot2`<sup>70</sup>.

The identification of protein classes corresponding to genes harboring an antisense transcript was performed with PANTHER tools (<https://pantherdb.org>).

## Data availability

The data from experiments of TSS identification and heat-treated amoebae (dataset 2 and 3 in this paper) are available in the SRA database (<https://www.ncbi.nlm.nih.gov/sra/>) under the accession number PRJNA615171. The data concerning virulence conditions (dataset 1) were already in the ArrayExpress database ([www.ebi.ac.uk/arrayexpress](http://www.ebi.ac.uk/arrayexpress)) under accession number E-MTAB-4882.

Received: 16 July 2020; Accepted: 2 November 2020

Published online: 19 November 2020

## References

1. Stensvold, C. R. *et al.* Increased sampling reveals novel lineages of *Entamoeba*: consequences of genetic diversity and host specificity for taxonomy and molecular detection. *Protist* **162**, 525–541. <https://doi.org/10.1016/j.protis.2010.11.002> (2011).
2. Marie, C. & Petri, W. A. Jr. Regulation of virulence of *Entamoeba histolytica*. *Annu. Rev. Microbiol.* **68**, 493–520. <https://doi.org/10.1146/annurev-micro-091313-103550> (2014).
3. Loftus, B. *et al.* The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868. <https://doi.org/10.1038/nature03291> (2005).
4. Lorenzi, H. A. *et al.* New assembly, reannotation and analysis of the *Entamoeba histolytica* genome reveal new genomic features and protein content information. *PLoS Negl. Trop. Dis.* **4**, e716. <https://doi.org/10.1371/journal.pntd.0000716> (2010).
5. Clark, C. G. *et al.* Structure and content of the *Entamoeba histolytica* genome. *Adv. Parasitol.* **65**, 51–190. [https://doi.org/10.1016/S0065-308X\(07\)65002-7](https://doi.org/10.1016/S0065-308X(07)65002-7) (2007).
6. Wilihoef, U., Campos-Gongora, E., Touzni, S., Bruchhaus, I. & Tannich, E. Introns of *Entamoeba histolytica* and *Entamoeba dispar*. *Protist* **152**, 149–156. <https://doi.org/10.1078/1434-4610-00053> (2001).
7. Davis, P. H., Schulze, J. & Stanley, S. L. Jr. Transcriptomic comparison of two *Entamoeba histolytica* strains with defined virulence phenotypes identifies new virulence factor candidates and key differences in the expression patterns of cysteine proteases, lectin light chains, and calmodulin. *Mol. Biochem. Parasitol.* **151**, 118–128. <https://doi.org/10.1016/j.molbiopara.2006.10.014> (2007).
8. Zamorano, A. *et al.* In silico analysis of EST and genomic sequences allowed the prediction of cis-regulatory elements for *Entamoeba histolytica* mRNA polyadenylation. *Comput. Biol. Chem.* **32**, 256–263. <https://doi.org/10.1016/j.compbiolchem.2008.03.019> (2008).
9. Hon, C. C. *et al.* Surveying *Entamoeba histolytica* transcriptome using massively parallel cDNA sequencing. In *Amebiasis* (eds Nozaki, T. & Bhattacharya, A.) 99–117 (Springer, Tokyo, 2015). [https://doi.org/10.1007/978-4-431-55200-0\\_7](https://doi.org/10.1007/978-4-431-55200-0_7).
10. Ehrenkauf, G. M., Hackney, J. A. & Singh, U. A developmentally regulated Myb domain protein regulates expression of a subset of stage-specific genes in *Entamoeba histolytica*. *Cell Microbiol.* **11**, 898–910. <https://doi.org/10.1111/j.1462-5822.2009.01300.x> (2009).
11. Santi-Rocca, J. *et al.* The lysine- and glutamic acid-rich protein KERP1 plays a role in *Entamoeba histolytica* liver abscess pathogenesis. *Cell Microbiol.* **10**, 202–217. <https://doi.org/10.1111/j.1462-5822.2007.01030.x> (2008).
12. Weber, C. *et al.* Extensive transcriptome analysis correlates the plasticity of *Entamoeba histolytica* pathogenesis to rapid phenotype changes depending on the environment. *Sci. Rep.* **6**, 35852. <https://doi.org/10.1038/srep35852> (2016).
13. Biller, L. *et al.* Differences in the transcriptome signatures of two genetically related *Entamoeba histolytica* cell lines derived from the same isolate with different pathogenic properties. *BMC Genomics* **11**, 63. <https://doi.org/10.1186/1471-2164-11-63> (2010).
14. Varet, H. *et al.* Enteric bacteria boost defences against oxidative stress in *Entamoeba histolytica*. *Sci. Rep.* **8**, 9042. <https://doi.org/10.1038/s41598-018-27086-w> (2018).
15. Vicente, J. B., Ehrenkauf, G. M., Saraiva, L. M., Teixeira, M. & Singh, U. *Entamoeba histolytica* modulates a complex repertoire of novel genes in response to oxidative and nitrosative stresses: implications for amebic pathogenesis. *Cell Microbiol.* **11**, 51–69. <https://doi.org/10.1111/j.1462-5822.2008.01236.x> (2009).
16. Hackney, J. A., Ehrenkauf, G. M. & Singh, U. Identification of putative transcriptional regulatory networks in *Entamoeba histolytica* using Bayesian inference. *Nucleic Acids Res.* **35**, 2141–2152. <https://doi.org/10.1093/nar/gkm028> (2007).
17. Gilchrist, C. A. *et al.* Regulation of Virulence of *Entamoeba histolytica* by the URE3-BP Transcription Factor. *mBio*. <https://doi.org/10.1128/mBio.00057-10> (2010).
18. Naiyer, S. *et al.* Transcriptomic analysis reveals novel downstream regulatory motifs and highly transcribed virulence factor genes of *Entamoeba histolytica*. *BMC Genomics* **20**, 206. <https://doi.org/10.1186/s12864-019-5570-z> (2019).
19. Hon, C. C. *et al.* Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*. *Nucleic Acids Res.* **41**, 1936–1952. <https://doi.org/10.1093/nar/gks1271> (2013).
20. Doolittle, W. F. We simply cannot go on being so vague about ‘function’. *Genome Biol.* **19**, 223. <https://doi.org/10.1186/s13059-018-1600-4> (2018).
21. Zhang, Y., Liu, X. S., Liu, Q. R. & Wei, L. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.* **34**, 3465–3475. <https://doi.org/10.1093/nar/gkl473> (2006).
22. Beiter, T., Reich, E., Williams, R. W. & Simon, P. Antisense transcription: a critical look in both directions. *Cell Mol. Life Sci.* **66**, 94–112. <https://doi.org/10.1007/s00018-008-8381-y> (2009).
23. Zhang, H., Ehrenkauf, 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.* **4**, e1000219. <https://doi.org/10.1371/journal.ppat.1000219> (2008).
24. Mar-Aguilar, F. *et al.* Identification and characterization of microRNAs from *Entamoeba histolytica* HM1-IMSS. *PLoS ONE* **8**, e68202. <https://doi.org/10.1371/journal.pone.0068202> (2013).
25. Saha, A., Bhattacharya, S. & Bhattacharya, A. Serum stress responsive gene EhslnRNA of *Entamoeba histolytica* is a novel long noncoding RNA. *Sci. Rep.* **6**, 27476. <https://doi.org/10.1038/srep27476> (2016).

26. Morrissy, A. S., Griffith, M. & Marra, M. A. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res.* **21**, 1203–1212. <https://doi.org/10.1101/gr.113431.110> (2011).
27. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255. <https://doi.org/10.1038/nature08756> (2010).
28. Gu, W. *et al.* CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**, 1488–500. <https://doi.org/10.1016/j.cell.2012.11.023> (2012).
29. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731. <https://doi.org/10.1038/nrg3293> (2012).
30. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335–338. <https://doi.org/10.1126/science.1181421> (2010).
31. Singh, U., Rogers, J. B., Mann, B. J. & Petri, W. A. Jr. Transcription initiation is controlled by three core promoter elements in the *hgl5* gene of the protozoan parasite *Entamoeba histolytica*. *Proc. Natl. Acad. Sci. USA* **94**, 8812–8817. <https://doi.org/10.1073/pnas.94.16.8812> (1997).
32. Bruchhaus, I., Leippe, M., Lioutas, C. & Tannich, E. Unusual gene organization in the protozoan parasite *Entamoeba histolytica*. *DNA Cell Biol.* **12**, 925–933. <https://doi.org/10.1089/dna.1993.12.925> (1993).
33. Dugar, G. *et al.* High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet.* **9**, e1003495. <https://doi.org/10.1371/journal.pgen.1003495> (2013).
34. Zhukova, A. *et al.* Genome-wide transcriptional start site mapping and sRNA identification in the pathogen. *Front. Cell Infect. Microbiol.* **7**, 10. <https://doi.org/10.3389/fcimb.2017.00010> (2017).
35. Purdy, J. E., Pho, L. T., Mann, B. J. & Petri, W. A. Upstream regulatory elements controlling expression of the *Entamoeba histolytica* lectin. *Mol. Biochem. Parasitol.* **78**, 91–103. [https://doi.org/10.1016/s0166-6851\(96\)02614-x](https://doi.org/10.1016/s0166-6851(96)02614-x) (1996).
36. Burke, T. W. & Kadonaga, J. T. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **10**, 711–724. <https://doi.org/10.1101/gad.10.6.711> (1996).
37. Wight, M. & Werner, A. The functions of natural antisense transcripts. *Essays Biochem.* **54**, 91–101. <https://doi.org/10.1042/bse0540091> (2013).
38. Georg, J. & Hess, W. R. Widespread antisense transcription in prokaryotes. *Microbiol. Spectr.* <https://doi.org/10.1128/microbiolspec.RWR-0029-2018> (2017).
39. Landry, S. *et al.* Detection, characterization and regulation of antisense transcripts in HIV-1. *Retrovirology* **4**, 71. <https://doi.org/10.1186/1742-4690-4-71> (2007).
40. Wagner, E. G., Altuvia, S. & Romby, P. Antisense RNAs in bacteria and their genetic elements. *Adv. Genet.* **46**, 361–398. [https://doi.org/10.1016/s0065-2660\(02\)46013-0](https://doi.org/10.1016/s0065-2660(02)46013-0) (2002).
41. Siegel, T. N. *et al.* Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. *BMC Genomics* **15**, 150. <https://doi.org/10.1186/1471-2164-15-150> (2014).
42. Steigele, S. & Nieselt, K. Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. *Nucleic Acids Res.* **33**, 5034–5044. <https://doi.org/10.1093/nar/gki804> (2005).
43. Mol, J. N. *et al.* Regulation of plant gene expression by antisense RNA. *FEBS Lett.* **268**, 427–430. [https://doi.org/10.1016/0014-5793\(90\)81298-3](https://doi.org/10.1016/0014-5793(90)81298-3) (1990).
44. Lee, S. *et al.* Detecting novel low-abundant transcripts in *Drosophila*. *RNA* **11**, 939–946. <https://doi.org/10.1261/rna.7239605> (2005).
45. Sun, M., Hurst, L. D., Carmichael, G. G. & Chen, J. Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity. *Genome Res.* **16**, 922–933. <https://doi.org/10.1101/gr.5210006> (2006).
46. Wanowska, E., Kubiak, M. R., Rosikiewicz, W., Makalowska, I. & Szczesniak, M. W. Natural antisense transcripts in diseases: From modes of action to targeted therapies. *Wiley Interdiscip. Rev. RNA* <https://doi.org/10.1002/wrna.1461> (2018).
47. Georg, J. & Hess, W. R. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 286–300. <https://doi.org/10.1128/MMBR.00032-10> (2011).
48. Lapidot, M. & Pilpel, Y. Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep.* **7**, 1216–1222. <https://doi.org/10.1038/sj.embor.7400857> (2006).
49. Rosengarten, R. D., Santhanam, B., Kokosar, J. & Shaulsky, G. The long noncoding RNA transcriptome of dictyostelium discoideum development. *G3 (Bethesda)* **7**, 387–398. <https://doi.org/10.1534/g3.116.037150> (2017).
50. Zheng, L. L. *et al.* Comparative transcriptome analysis of small noncoding RNAs in different stages of *Trypanosoma brucei*. *RNA* **19**, 863–875. <https://doi.org/10.1261/rna.035683.112> (2013).
51. Thomason, M. K. & Storz, G. Bacterial antisense RNAs: how many are there, and what are they doing?. *Annu. Rev. Genet.* **44**, 167–188. <https://doi.org/10.1146/annurev-genet-102209-163523> (2010).
52. Lavorgna, G. *et al.* In search of antisense. *Trends Biochem. Sci.* **29**, 88–94. <https://doi.org/10.1016/j.tibs.2003.12.002> (2004).
53. Osato, N., Suzuki, Y., Ikeo, K. & Gojobori, T. Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* **176**, 1299–1306. <https://doi.org/10.1534/genetics.106.069484> (2007).
54. Lloréns-Rico, V. *et al.* Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv* **2**, e1501363. <https://doi.org/10.1126/sciadv.1501363> (2016).
55. Park, H., Yoon, Y., Suk, S., Lee, J. Y. & Lee, Y. Effects of different target sites on antisense RNA-mediated regulation of gene expression. *BMB Rep.* **47**, 619–624 (2014).
56. Criscuolo, A. & Brisse, S. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102**, 500–506. <https://doi.org/10.1016/j.ygeno.2013.07.011> (2013).
57. Aurrecochea, C. *et al.* AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res.* **39**, D612–D619. <https://doi.org/10.1093/nar/gkq1006> (2011).
58. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr. Protoc. Bioinformatics.* <https://doi.org/10.1002/0471250953.bi1114s51> (2015).
59. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
60. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. <https://doi.org/10.1038/nbt.1883> (2011).
61. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875. <https://doi.org/10.1093/bioinformatics/bti310> (2005).
62. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930. <https://doi.org/10.1093/bioinformatics/btt656> (2014).
63. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25. <https://doi.org/10.1186/gb-2009-10-3-r25> (2009).
64. Quinlan, A. R. BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinf.* <https://doi.org/10.1002/0471250953.bi1112s47> (2014).
65. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf.* **11**, 94. <https://doi.org/10.1186/1471-2105-11-94> (2010).

66. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659. <https://doi.org/10.1093/bioinformatics/btr261> (2011).
67. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. <https://doi.org/10.1186/s13059-014-0550-8> (2014).
68. 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* **11**, e0157022. <https://doi.org/10.1371/journal.pone.0157022> (2016).
69. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
70. Wickham, H. *et al.* *ggplot2: elegant graphics for data analysis* (Springer-Verlag, New York, 2016).
71. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. <https://doi.org/10.1038/nbt.1754> (2011).

## Acknowledgements

The authors gratefully acknowledge the help of Hugo Varet (Hub de Bioinformatique et Biostatistique – DBC, Institut Pasteur) for his advice on differential expression experiments, to Pascal Campagne (Hub de Bioinformatique et Biostatistique – DBC, Institut Pasteur) for critical reading of the manuscript and to Susanne Warrenfeltz (AmoebaDB) for helping us in the search for gene pairs.

## Author contributions

C.W. and M.K. 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., J.Y.C. and N.G. wrote the manuscript.

## Funding

The work was funded by the National French Research Agency (ANR-2010-GENM-0011-01, GENAMIBE). The Transcriptome and Epigenome Platform is a member of the France Génomique consortium (ANR10-NBS-09-08). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-77010-4>.

**Correspondence** and requests for materials should be addressed to D.M. or N.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



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

© The Author(s) 2020