1936–1952 *Nucleic Acids Research, 2013, Vol. 41, No. 3* *Published online 20 December 2012*
doi:10.1093/nar/gks1271

Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*

Chung-Chau Hon[1,2,*], Christian Weber[1,2], Odile Sismeiro[3], Caroline Proux[3], Mikael Koutero[3], Marc Deloger[1,2], Sarbashis Das[5], Mridula Agrahari[4], Marie-Agnes Dillies[3], Bernd Jagla[3], Jean-Yves Coppee[3], Alok Bhattacharya[4,5] and Nancy Guillen[1,2]

[1]Institut Pasteur, Unité Biologie Cellulaire du Parasitisme, Département Biologie cellulaire et infection, F-75015 Paris, France, [2]INSERM U786, F-75015 Paris, France, [3]Institut Pasteur, Plate-forme Transcriptome et Epigénome, Département Génomes et Génétique, F-75015 Paris, France, [4]Jawaharlal Nehru University, School of Life Sciences, New Delhi 110067, India, and [5]Jawaharlal Nehru University, School of Computational and Integrative Sciences, New Delhi 110067, India

Received July 17, 2012; Revised November 3, 2012; Accepted November 5, 2012

ABSTRACT

Alternative splicing and polyadenylation were observed pervasively in eukaryotic messenger RNAs. These alternative isoforms could either be consequences of physiological regulation or stochastic noise of RNA processing. To quantify the extent of stochastic noise in splicing and polyadenylation, we analyzed the alternative usage of splicing and polyadenylation sites in *Entamoeba histolytica* using RNA-Seq. First, we identified a large number of rarely spliced alternative junctions and then showed that the occurrence of these alternative splicing events is correlated with splicing site sequence, occurrence of constitutive splicing events and messenger RNA abundance. Our results implied the majority of these alternative splicing events are likely to be stochastic error of splicing machineries, and we estimated the corresponding error rates. Second, we observed extensive microheterogeneity of polyadenylation cleavage sites, and the extent of such microheterogeneity is correlated with the occurrence of constitutive cleavage events, suggesting most of such microheterogeneity is likely to be stochastic. Overall, we only observed a small fraction of alternative splicing and polyadenylation isoforms that are unlikely to be solely stochastic, implying the functional relevance of alternative splicing and polyadenylation in *E. histolytica* is limited. Lastly, we revised the gene models and annotated their 3′UTR in AmoebaDB, providing valuable resources to the community.

INTRODUCTION

In higher eukaryotes, alternative splicing and polyadenylation, which generate multiple isoforms from a single messenger RNA (mRNA) precursor (pre-mRNA), are the major mechanisms for expanding the diversity of their transcriptomes and proteomes (1,2). Numerous studies demonstrated alternative splicing is pervasive in higher eukaryotes. For example, ~95% of multi-exon genes undergo alternative splicing in human (3,4) and at least 42% of intron-containing genes are alternatively spliced in *Arabidopsis thaliana* (5). Moreover, microheterogeneity (6) and long-range heterogeneity (7) of polyadenylation site usage in eukaryotic mRNAs are also found to be extensive. Recent studies of polyadenylation site heterogeneity using RNA-Seq further demonstrated the pervasiveness of alternative polyadenylation in animals (2,8,9) and plants (10,11). Functional consequences of physiologically regulated alternative splicing are well documented (12). Also, the impacts of alternative polyadenylation on mRNA coding capacity, localization, translation efficiency and stability have also been described (13). Nonetheless, the proportion of these alternative isoforms being physiologically regulated versus those that are solely derived from the inherent stochasticity of RNA processing (14) is largely unknown.

How much of the observed alternative splicing and polyadenylation are a consequence of stochastic noise of RNA processing? A number of studies attempted to address this question. Based on comparative analyses of human and mouse expressed sequence tags (EST) data, Sorek *et al.* (15) proposed a significant portion of alternative isoforms is likely to be non-functional, and might be resulted from aberrant rather than regulated splicing

*To whom correspondence should be addressed. Tel: +33 1 45 68 86 75; Fax: +33 1 45 68 86 74; Email: chung-chau.hon@pasteur.fr

events. Melamud and Moult (14) further showed that the number of alternative isoforms and their abundance can be predicted by a simple stochastic noise model, demonstrating most alternative splicing in humans is a consequence of stochastic noise in the splicing machineries. More recently, Pickrell *et al.* (16) used RNA-Seq to demonstrate the existence of a large class of low abundance and unconserved isoforms, presenting empirical data to support the hypothesis of noisy splicing. The extent of stochastic noise in polyadenylation is less well studied, despite this the genome-wide atlas of polyadenylation site was mapped in a number of model organisms (2,9,11). Quantifying the properties of alternative splicing and polyadenylation events in wider range of eukaryotes would certainly help to clarify the inherent stochasticity of these processes, and hence provide insight into the prevalence of functionally relevant alternative isoforms.

In this study, we sequenced the poly(A)+ transcriptome of *Entamoeba histolytica* at saturated depth and quantified the extent of alternative usage of splicing and polyadenylation sites in its mRNAs. *E. histolytica* is an enteric parasite in humans, which causes amoebiasis in ~10% of the infected individuals, resulting in 50 million cases of dysentery annually (17). *Entamoeba* belongs to the Amoebozoa kingdom, which represent one of the earliest branches from the last common ancestor of all eukaryotes and is phylogenetically distinct from 'model organisms' of animals, fungi and plants (18). While most of the observations on alternative splicing and polyadenylation were derived from studying these model organisms of animals, fungi and plants, it is therefore interesting to extend the observations to other less characterized kingdoms.

Initial analyses of *E. histolytica* genome in 2005 (assembly of ~23 Mb with 888 scaffolds) predicted 9938 coding genes (average size: 1.17 kb), comprising 49% region of the genome (19). About 25% of these genes were predicted to contain introns, and only 6% of them contain multiple introns (19). This initial analysis provided the first blueprint of *E. histolytica* genome to the research community, which opened the avenue to post-genomic high-throughput studies, e.g. transcriptomics and proteomics. Nonetheless, the genome is AT rich and highly repetitive, and thus, this initial assembly might contain misassembled regions and partially sequenced or unidentified genes (20). Therefore, the genome was reassembled 5 years after its initial analyses, with >100 artifactual tandem duplications eliminated, reducing the assembly size to ~20 Mb with 1496 scaffolds (21). Re-annotation of the new assembly reduced the predicted gene number to 8201, and 40% of the original gene models were changed (21). Even so, most of the gene models were solely based on *in silico* prediction and lack of supporting experimental data, e.g. complementary DNA (cDNA)/EST.

The primary goal of this study is to quantify the heterogeneity of splicing and polyadenylation in *E. histolytica*, an organism with few introns and short 3′ untranslated region (UTR) (22,23), providing insights into the stochastic noise of these processes in lower eukaryotes. In addition, as resources for the *Entamoeba* community, we revised the gene model annotations of *E. histolytica* in AmoebaDB based on our sequencing data.

MATERIALS AND METHODS

E. histolytica strains, RNA extraction, construction and sequencing of cDNA libraries

Two well-characterized strains of *E. histolytica*, HM1:IMSS and Rahman, were cultivated axenically in TYI-S-33 medium at 37°C, with three biological replicates for each. Trophozoites in log phase of growth were collected. HM1:IMSS is a prototype virulent strain, and Rahman is an avirulent strain (24). Total RNA was extracted from each of the biological replicates ($n = 6$) using Trizol reagents, and poly(A)+ mRNA was purified from total RNA using Dynabeads according manufacturer's instructions (Invitrogen). For high-throughput sequencing, paired-end cDNA libraries ($n = 6$) were prepared from poly(A)+ mRNA according to manufacturer's instructions (mRNA-Seq 8-Sample Prep Kit, Illumina). cDNA fragments of ~200 bp were purified from each library and were sequenced from both ends for 100 bp, using an Illumina HiSeq2000 instrument according manufacturer's instructions (Illumina). The RNA-Seq data have been deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena/data/view/ERP001024).

Reference genome and gene annotations

Genome scaffolds of *E. histolytica* HM1:IMSS version 1.3 in AmoebaDB were used as the reference genome for all analyses (25). Gene models were retrieved from AmoebaDB version 1.3; repetitive elements were defined by previously annotations (26).

Read mapping and splicing junction identification

Reads with longer than 50 nt flagged with Illumina's low quality flag 'B' were removed from the data sets. The reads were mapped, as pairs, onto the reference genome using Bowtie version 0.12.7 (27) with the following parameters: (i) maximum two mismatches were allowed in a 50-nt seed region (i.e. −n 2 −l 50); and (ii) reads mapped to >50 locations were discarded (i.e. −m 50), and reads mapped to multiple locations were reported only once (i.e. −k 1). To identify the splicing junctions, the unaligned reads from Bowtie were mapped using HMMSplicer version 0.95 (28) with following parameters: (i) maximum two mismatches in second half matching (i.e. −e 2); (ii) maximum intron size of 1000 nt (i.e. −k 1000); and (iii) anchor size of 10 nt (i.e. −a 10). To be conservative, we have chosen HMMSplicer cutoff scores of 700 for a single read (default = 600) and 500 for multiple reads (default = 400), which are both higher than the default values.

Definition of splicing efficiency, junction clusters and constitutive isoforms

First, we pooled the mapping results of all libraries to define transcribed fragments (i.e. transfrags). A transfrag is defined as a continuous genomic region of ≥100 nt that is covered by ≥3 reads per nucleotide with gap size ≤20 nt.

Junctions identified in all libraries were also pooled. Splicing efficiency of a junction on an mRNA is defined as the ratio of the number of reads supporting this junction to the coverage (per nucleotide) of the corresponding mRNA. By definition, ideally, splicing efficiency of 1 implies 100% of the pre-mRNA molecules were spliced at this particular junction. We collapsed the junctions into 'clusters' as follows. Junctions spanning across multiple non-overlapping junctions were defined as 'super-junctions' (which involve exon skipping), and the rest were defined as 'infer-junctions'. Then, super- and infer-junctions were subjected to 'clustering' separately. Clustering refers to collapsing of overlapping junctions into clusters, and each cluster is represented by a 'representative junction', which has the highest number of supporting reads. Each infer-junction cluster was then classified either as 'constitutive' or 'alternative'. A junction (and thus the cluster it belongs to) is said to be 'constitutive' (or otherwise 'alternative') if the ratio of its exonic flanking coverage to intronic flanking coverage >0.3, i.e. at least one-third of all reads covering the junction region, is spliced (explained in detailed in Supplementary Figure S1). Constitutive junctions hence refer to junctions that are constitutively spliced on a given pre-mRNA. Constitutive isoform of a gene was constructed based on the combinations of constitutive junction clusters and overlapping transfrags. These constitutive isoforms were then used to revise the gene models. Please see Supplementary Figure S2 for a detailed workflow.

Correlating splicing efficiency to the degree of sequence conservation of splicing sites

We quantified the degree of sequence conservation of splicing sites and investigate its correlation with splicing efficiency. We used 'consensus value (CV)', adopted from Shapiro and Senapathy (29) with slight modifications, which reflects the sequence similarity of a splicing site to the consensus of a reference set of splicing sites. For the reference set of splicing site, we used the top 5% ($n = 242$) splicing junctions with highest splicing efficiency. For splicing site sequences, we refer to the 20-mers of intronic sequences next to the 5′ and 3′ splicing sites excluding the GU-AG site, i.e. position +3 to +12 of 5′ splicing sites and −12 to −3 of 3′ splicing sites. Then, we aligned these sequences and established a reference position-specific scoring matrix (PSSM) based on the alignment. We calculate the CV of each splicing site as:

$$CV = \left(\frac{\sum_i f_{i,n} - \sum_i f_{i,min}}{\sum_i f_{i,\max} - \sum_i f_{i,min}} \right),$$

where i is the position of a nucleotide along the 20-mer mentioned earlier; n is the actual nucleotide sequence at position i; $f_{i,\,n}$ is the relative frequency of nucleotide n in the reference PSSM; $f_{i,\,\min}$ is the relative frequency of the rarest nucleotide at position i in the reference PSSM; $f_{i,\,\max}$ is the relative frequency of most common nucleotide at position i in the reference PSSM. Therefore, CV is a number ranging from 0 to 1, quantifying the sequence conservation between a single splicing site and a group of most efficiently spliced sites (i.e. reference PSSM). To investigate the correlation between CV and splicing efficiency, splicing site sequences of discrete splicing efficiency intervals (in $\log_{10}$ scale, bin width = 0.1) were pooled, and the mean CV were plotted against the mean splicing efficiency within each interval. Sequence logos and conservation at particular positions (i.e. measured in entropy 'bits') were calculated using Weblogo (30).

Discovering the motifs for splicing sites of alternative intron creation events

Enrichment of sequence motifs around the alternatively created introns was tested using discriminative regular expression motif elicitation (DREME) (31). DREME performs *discriminative* discovery of motif that is enriched in a positive set in comparison with a negative set. First, intronic (20 nt) and exonic (50 nt) sequences flanking the splicing sites of alternative intron creation junctions were extracted. Alternative intron creation junctions were defined as non-stochastic ($n = 137$, as positive set) and stochastic ($n = 332$, as negative set) when its splicing efficiency is ≥ 0.13 and ≤ 0.013, respectively. Flanking sequences of constitutive splicing sites ($n = 2269$) and coding sequences randomly sampled from non-intronic genes ($n = 5000$) were used as control positive set and negative set, respectively. Enrichment of motifs of ≥ 4 nt and ≤ 10 nt was tested in the positive sets against the negative sets using DREME (31).

Gene model revision

We developed a Perl script to scan for the intersections between the constitutive isoform (from our RNA-Seq data set) and existing gene models (from AmoebaDB), and automatically categorized the conflicts between them. These conflicts were then revised manually to define a set of *bona fide* gene models that satisfies the following criteria. Inherent criterion: it contains a complete open reading frame (ORF) and is located at least 100 nt away from ambiguous regions of the scaffold (i.e. scaffold ends and regions with 'Ns'); coverage criterion: >95% area of its ORF is covered by at least one read in the pooled data; junction criterion: all of its junctions (if any) are validated in the pooled data set. A novel transcript was conservatively defined as transfrags located at least 100 nt away from existing annotations and ambiguous regions. Please see Supplementary Figure S3 for detailed workflow.

Identifying the polyadenylation sites from reads

The criteria used here are primarily based on Lee *et al.* (32). Briefly, reads containing five or more consecutive 'A' at their end (or 'T' at their beginning, which will be reverse complemented in downstream analyses) were selected from each of the six libraries, and redundant reads were removed. These non-redundant reads were pooled. These reads potentially contain the sequence of mRNA poly(A) tails. The A stretch at the end were trimmed, and the reads with minimum 18 nt after trimming were mapped to the reference genome using Bowtie with parameters '−n 2 −k 1 −m 50 −l 30'. To distinguish poly(A) tracks of

true polyadenylation from poly(A) tracks of internal poly(A) stretches on the mRNAs themselves (i.e. false positives), we analyzed the base compositions surrounding the end of the mapped reads and discard those that might not represent true polyadenylation. Reads with the following properties were regarded as false positives and removed. (i) Reads with ≥5 nt immediate downstream of the end site are A's; (ii) depending on the actual length of the poly(A) stretch of the read (e.g. N nt), reads with 70% of N nt downstream of the end site are A's; and (iii) reads with ≥8 nt within 10 nt immediate upstream of the end site are A's. The polyadenylation sites were then defined as the immediate downstream base of the reads. To ensure the identified polyadenylation sites are not false positives derived from low quality base calls, reads with quality scores in any of the upstream and downstream 5 nt flanking the polyadenylation site <20 were further removed. These procedures should be able to remove false positives derived from internal poly(A) stretches and low quality base calls.

Assigning the poly(A) site clusters to gene models

As most of the observed polyadenylation sites appear as clusters (6), we grouped the poly(A) sites into clusters by allowing an optimal maximum intra-cluster distance (at 12 nt) between sites (Figure 7A). A poly(A) cluster was then represented by the poly(A) site with highest number of supporting reads (i.e. peak), and these peak positions were used in all downstream analyses. A poly(A) cluster is defined as valid when the number of reads at the peak position is ≥2. To assign poly(A) tails to mRNAs, we searched for poly(A) clusters within 200 nt downstream of their stop codons on the same strand and recorded the size of the coverage gap between the poly(A) clusters and the stop codon. A poly(A) tail is defined as valid when coverage gap is ≤10 nt. Length of 3′UTR of an mRNA is defined as distance between the nucleotide (inclusive) after its stop codon and the nucleotide (inclusive) before peak site of the farthest valid poly(A) clusters.

Definition of microheterogeneity of polyadenylation and long-range alternative polyadenylation

Microheterogeneity of polyadenylation is defined as the alternative poly(A) sites within a poly(A) cluster. Long-range alternative polyadenylation in mRNAs, including multiple distinct poly(A) clusters in 3′UTR and poly(A) clusters within ORF, were defined according to the following criteria. To be conservative, we only consider poly(A) clusters having ≥3 supporting reads with average poly(A) track length ≥10 nt. For multiple distinct poly(A) clusters in 3′UTR, the clusters have to be at least 25 nt away from each other and within 100 nt from stop codons of corresponding mRNAs. A minimum inter-cluster distance of 25 nt was chosen because >98% of poly(A) clusters are sized ≤25 nt (Figure 7B). A maximum distance of 100 nt from stop codon was chosen because >97% of all poly(A) sites are within 100 nt from their corresponding stop codon. For poly(A) clusters within ORF, the clusters have to be ≤25 nt upstream to the corresponding stop codons, and the number of A's in the 10 nt upstream and downstream of the peak site has to be ≤4. A cutoff of minimum 25 nt upstream to stop codon was chosen to avoid the possibility that the poly(A) clusters within ORF were actually part of the poly(A) clusters at 3′UTR. A cutoff of maximum four As in the 10 nt flanking the peak site was chosen to avoid the possibility that the A-tracks on poly(A) reads were artifactual because of A-rich regions on mRNA.

Discovering the sequences motifs for polyadenylation

The sequences immediate upstream and downstream (50 nt on each side) of the poly(A) site of all mRNAs (n = 5018) were used to scan for conserved motifs using DREME (31). 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. Highly stringent E-value (10^{-50}) was chosen to avoid spurious motifs. 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 (200 nt) the poly(A) sites.

Counting the occurrence of AAWUDA motifs

A motif, AAWUDA, was found significantly enriched at around 20 nt upstream of all poly(A) clusters at 3′UTR (see 'Results' section). To investigate the occurrence of this motif in long-range alternative poly(A) clusters in 3′UTR (n = 102) and within ORFs (n = 59), we extracted 15 nt from position −25 to −10 nt of their peak sites and counted the occurrence of the motif and the proportion of its variants (i.e. AAUUAA, AAUUUA, AAAUAA, AAAUUA, AAUUGA and AAAUGA). To assess the background occurrence of this motif in the target sequences, we counted both real sequences and permutated sequences (i.e. the 15 nt real sequence being randomly shuffled). As positive and negative controls, sequences extracted from 3′UTRs with a single poly(A) cluster (n = 4173) and random genomic sequences of the same length (n = 10 000) were also counted, respectively.

Identifying differentially expressed alternative splicing isoforms between HM1:IMSS and Rahman

Non-stochastic splicing isoforms (n = 194) were defined as mentioned in the 'Results' section, and their genomic coordinates were combined with that of the reference gene models (n = 7312, revised gene models) into a single annotation file in general transfer format (GTF). We then remapped the reads of the six libraries separately onto these reference and alternative isoforms using TopHat version 2.0.4 (33) by supplying the mentioned GTF file, and *de novo* identification of splicing junction was disabled. Default settings were used for other parameters. Outputs from TopHat were then analyzed using Cufflinks package version 2.0.2 (34) for identification of differentially expressed alternative splicing isoforms. We also counted the number of reads supporting these alternative junctions between the two strains. If the read number of a junction is ≥5 in both strains and the fold

change of read number between the two strains is ≥3, the alternative isoform derived from this junction is said to express at substantially different levels. To assess the reproducibility of the biological replicates, we performed heuristic clustering and principle component analysis on their expression profiles (i.e. fragment per kilobase per millions values of all *bona fide* genes estimated in Cufflinks) using CummeRbund version 1.99.1 (http://compbio.mit.edu/cummeRbund/) and FactoMiner (35), respectively.

Identifying variability in alternative poly(A) clusters between HM1:IMSS and Rahman

For alternative poly(A) clusters, we considered only the multiple poly(A) clusters in 3′ UTR ($n = 102$) and poly(A) clusters within ORF ($n = 59$) as defined in the 'Results' section. For each of these alternative poly(A) clusters in each of the strains, we calculated the proportion of the alternative poly(A) events in the total poly(A) events of the corresponding genes (denoted *prptnAlt*) as the following:

$$prptnAlt = \frac{\text{Reads in the alternative poly(A) cluster}}{\text{Reads in all poly(A) clusters of the gene}}$$

To be conservative, we considered only alternative poly(A) clusters with at least 10 reads in the pooled data set of six libraries. The proportion of an alternative poly(A) cluster is said to be 'substantially shifted' between two strains when the fold change of *prptnAlt* between two strains is ≥2.

Analyses of gene ontology (GO)

We investigated the functional themes of the non-stochastic alternatively spliced genes and genes with long-range alternative polyadenylation sites using GO::TermFinder, which visualizes the gene ontology (GO) of a list of genes and performs enrichment tests (36). GO annotation of *E. histolytica* was retrieved from AmoebaDB. A threshold of *P*-value ≤0.05 and a false discovery rate ≤10% was chosen to define significant enrichment.

RESULTS

Data set coverage

In this study, we sequenced the poly(A)+ transcriptome of the prototype virulent strain, HM1:IMSS, and the avirulent strain, Rahman (24), with three biological replicates each ($n = 6$). The six data sets consist of ∼1000 million reads of 100 nt in total, covering the ∼20 Mb genome at ∼5000 times. Details of the data set were summarized in Supplementary Table S1. First, the six data sets were independently mapped to the genome (see 'Materials and Methods' section). Then, we assessed the reproducibility of the biological replicates. Both principle component analyses and heuristic clustering of their expression profiles (see 'Materials and Methods' section) suggest the replicates are generally reproducible, and the expression profiles from the two strains are readily distinguishable from each other (Supplementary Figure S4). For the purposes of gene model revision, as well as splicing junctions and polyadenylation sites identification, we pooled the mapping results (i.e. junctions and read pileup coverage) from the six data sets and evaluated its coverage on the annotated gene models. More than 98% of the annotated coding region was covered by at least two reads (Figure 1A), implying our data set is saturated and deep enough to cover most of annotated coding transcripts. Then, we quantified the portions of genome being transcribed as poly(A)+ RNA (Figure 1B). Here, we focus on the 'non-repetitive and non-structural RNA' genome, i.e. genomic regions excluding ribosomal RNA, transfer RNA and repetitive elements (Figure 1C, left panel). About 86% of the 'non-repetitive and non-structural RNA' genome was covered by at least two reads (Figure 1B), implying at least 86% of the 'non-repetitive and non-structural RNA' genome is transcribed (Figure 1C, right panel). We also noticed ∼24% transcribed area is unannotated (Figure 1C, right panel), reflecting the incompleteness of the current gene model annotation. Therefore, we identified novel coding transcripts and revised the existing gene models, defining a set of *bona fide* gene models ($n = 7312$), which will be described in the last section. It should be noted that 'mRNAs' and 'gene models' discussed in the following sections refer to these *bona fide* gene models.

Existence of a large number of unannotated junctions

To assess the extent of alternative splicing, we systematically analyzed the splicing junctions on a genome-wide scale. Here, we confine our discussion to canonical junctions mapped within the ORF of mRNAs on the same strand ($n = 6417$). First, 2089 of 2557 (∼81%) junctions annotated in AmoebaDB were confirmed by our data set, i.e. 468 of them were left unconfirmed (Supplementary Figure S5A). These unconfirmed junctions might be either because of (i) our data coverage is not wide and deep enough; or (ii) these junctions do not exist in reality (i.e. intron prediction errors). A previous study suggested that skewed ratio of intron length categories of 3n, 3n + 1 and 3n + 2 could be indicators of intron prediction errors, as intron lengths are not expected to respect coding frame (37). We thus compared the intron length category ratio of the confirmed and unconfirmed AmoebaDB junctions. In unconfirmed junctions, both 3n + 1 (∼36%) and 3n + 2 (∼38%) are substantially more frequent than 3n introns (∼25%), whereas in the confirmed junctions, the ratio of 3n + 1 (∼34%), 3n + 2 (∼35%) and 3n introns (∼31%) is mostly unbiased (Supplementary Figure S5B). For instance, ∼50% of these unconfirmed junctions are located either in proximity to the ambiguous genomic regions ($n = 154$), or within repetitive regions ($n = 77$), which are prone to assembly errors (Supplementary Figure S5A). We manually inspected some of these unconfirmed junctions located close to contig breaks and observed frame-shifting (i.e. 3n + 1 or 3n + 2) artifactual introns were often predicted to 'rescue' a false stop codon for the ORF (explained in Supplementary Figure S5C). These data suggest a substantial proportion of these unconfirmed junctions might be errors of intron prediction and

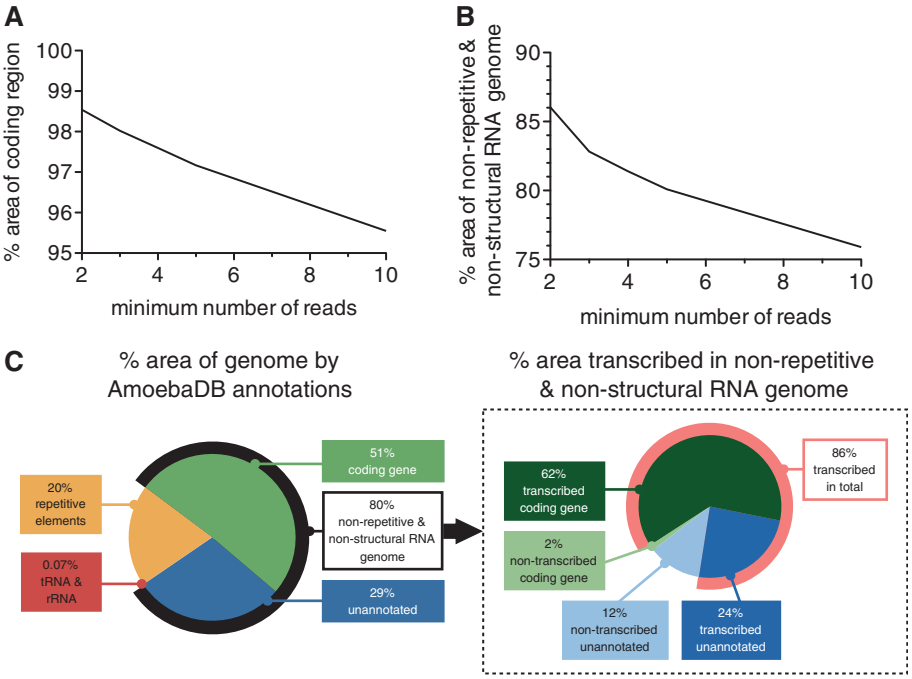

Figure 1. Data set coverage and proportion of transcribed genomic region. (**A**) Coverage of the annotated coding regions at various depths. The data suggest our data set covered most of the coding regions. (**B**) Coverage of 'non-repetitive and non-structural RNA' genome at various depths. (**C**) Left panel: composition of annotated genomic regions. The 'non-repetitive and non-structural RNA' genome refers to the genome excluding ribosomal RNAs, transfer RNAs and repetitive elements. Right panel: proportion of transcribed area of the 'non-repetitive and non-structural RNA' genome. About 24% of the 'non-repetitive and non-structural RNA' genome is transcribed but unannotated.

do not exist. In fact, 30% of unconfirmed junctions (n = 163) were corrected and led to modifications of gene models (discussed later). Notably, only 74 of them, i.e. <3% of all junctions annotated in AmoebaDB, could neither be confirmed nor explained as earlier, implying the depth of our data set should be enough to cover most, if not all, of the splicing junctions. Interestingly, the number of junctions identified here (n = 6417) substantially outnumbered the confirmed AmoebaDB junctions (n = 2089), suggesting the existence of a large number of unannotated junctions on mRNAs (n = 4328).

Most of the unannotated junctions are rarely spliced

To gain more insights into the aforementioned unannotated junctions, we investigated their splicing efficiency. We define 'splicing efficiency' as the ratio between the number of reads supporting a junction to the coverage (per nucleotide) of the corresponding mRNA (see 'Materials and Methods' section). The median splicing efficiency of junctions annotated in AmoebaDB (n = 2089) and those that are not annotated in AmoebaDB (n = 4328) are ~0.46 and ~0.0012, respectively (Figure 2). It implies most of the unannotated junctions are rarely spliced, comparing with the junctions annotated in AmoebaDB that are often spliced. Then, we used the 5% percentile of the splicing efficiency of junctions annotated in AmoebaDB (~0.03; Figure 2, dotted line) as the cutoff to define whether a junction is rarely or often spliced. Often-spliced junctions refer to junctions that are frequently, or constitutively, spliced off from the pre-mRNA molecules, whereas rarely-spliced junctions refer to the opposite scenario. There are 2262

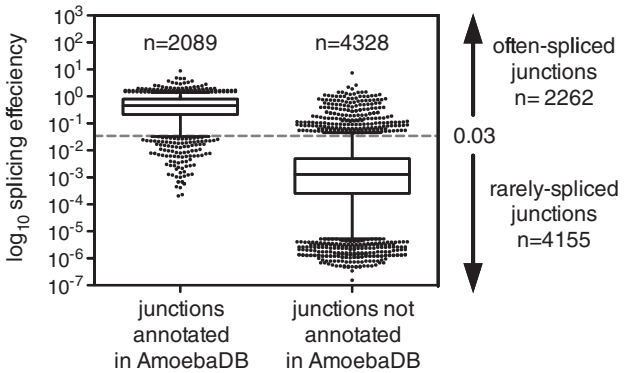

Figure 2. Defining often- and rarely-spliced junctions by splicing efficiency. The 'boxes and whiskers' represent the 5th, 25th, 50th, 75th and 95th percentiles. The dots represent data points below and above the 5th and 95th percentiles. The cutoff for defining often- and rarely-spliced is indicated by a dotted line.

often-spliced junctions, which, by definition, include 95% of the junctions annotated in AmoebaDB, and 227 unannotated junctions. On the other hand, there are 4155 rarely-spliced junctions, which are mostly unannotated. These results indicate most of the unannotated junctions are rarely spliced.

Splicing site sequence of the rarely-spliced junctions are less conserved

We reasoned if the majority of these rarely-spliced junctions were functionally relevant, their splicing efficiency should be regulated through physiological mechanisms

rather than determined by splicing site sequence. Alternatively, if majority of the rarely-spliced splicing junctions were a consequence of stochastic noise of splicing machineries, their splicing efficiency should be dependent on splicing site sequences, assuming certain composition of splicing site sequence was inherently more preferred by the splicing machineries. To this end, we investigated the correlation between splicing efficiency and sequence conservation of all junctions. First, the often-spliced junctions showed conserved splicing site sequences of 5′-GUUUGUU-UAG-3′ (Figure 3A), agreeing with the reported sequence in a previous study (23). However, the rarely-spliced junctions are generally less conserved (Figure 3B), suggesting the splicing efficiency of a junction might be correlated with its splicing site sequence. To further investigate such correlation, we pooled the sequences of splicing sites within discrete intervals of splicing efficiency and plotted the degree of sequence conservation within each interval (Figure 3C). As expected, sequence conservation (measured in entropy 'bits') of the 5′-GUUUGUU-UAG-3′ motif decreases with splicing efficiency (Figure 3C). To quantify such correlation, we plotted the mean CV against the mean splicing efficiency of each interval (Figure 3D). CV is a measurement of sequence similarity of a splicing site to the consensus of most efficiently spliced junctions (see 'Materials and Methods' section). The mean CV of splicing sites is significantly correlated with their mean splicing efficiency (Pearson's correlation coefficient = 0.97, $P < 0.0001$, 95% CI 0.92–0.99). It suggests the splicing sites of less efficiently spliced junctions are diverged from that of the most efficiently spliced junctions, and such divergence is quantitatively correlated with splicing efficiency. These results demonstrated the observed splicing efficiency of most junctions is dependent on composition of sequence motifs around the splicing site. It implies the splicing machineries in *Entamoeba* have a preference for specific splicing site sequence, and such preference seems to be the major factor affecting the splicing efficiency of most junctions. Taken together, these data strongly suggest the splicing efficiency of majority of the rarely-spliced junctions is determined by splicing site sequence rather than regulated through physiological mechanisms, implying they are likely to be a consequence of stochasticity of splicing machineries. In fact, ~69% of the rarely-spliced junctions overlap with at least one often-spliced junction (i.e. appear as 'junction clusters' as in Figure 4A); such association implies those often and 'constitutive' splicing events might be the source of these rare and 'alternative' splicing events. In the following sections, we defined the constitutive and alternative splicing events and investigated the correlation between their occurrences.

Definition of constitutive and alternative splicing events

To this end, we reconstructed the constitutive isoforms of all mRNAs and defined three types of alternative splicing events, including alternative 5′ or 3′ splicing, exon skipping and intron creation. First, we collapsed all overlapping junctions as clusters (Figure 4A), and categorize them as 'constitutive' or 'alternative' junction clusters, based on the ratio of exonic to intronic coverage surrounding the junction (explained in Supplementary Figure S1). A constitutive junction cluster, by definition, implies its intronic regions are constitutively being spliced off from the pre-mRNA. Alternative 5′ or 3′ splicing events were defined as alternative junctions within a constitutive junction cluster ($n = 1647$; Figure 4A). Exon skipping events were defined as alternative junctions that span across multiple constitutive junction clusters ($n = 99$; Supplementary Figure S1). Intron creation events were defined as alternative junctions that do not overlap with any constitutive junction cluster ($n = 1667$; Supplementary

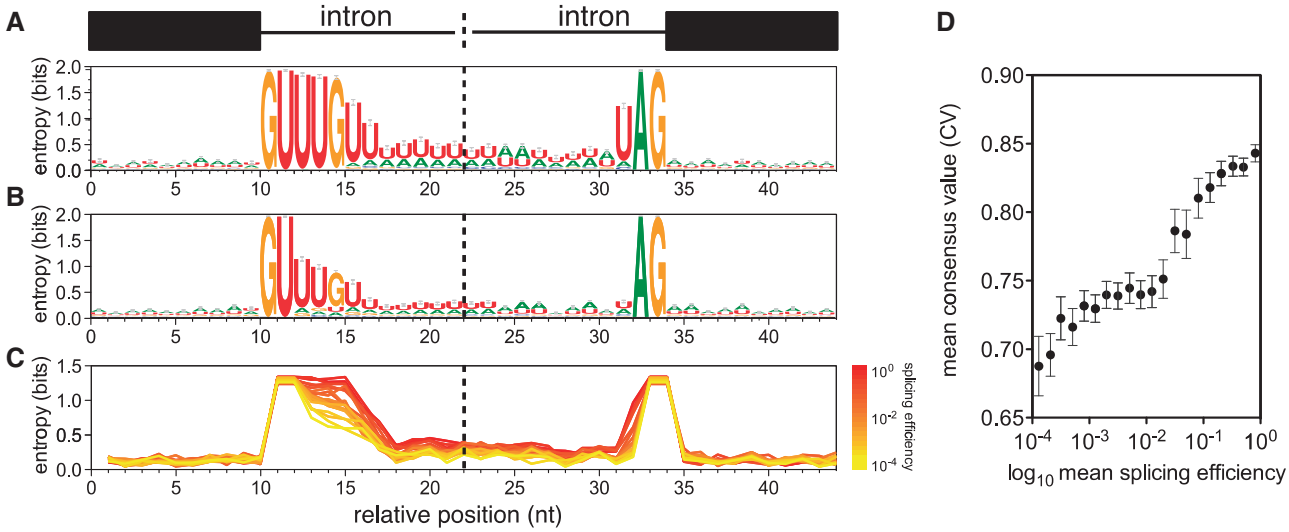

Figure 3. Splicing efficiency is correlated with splicing site sequence conservation. (**A**) Weblogo of often-spliced junctions. (**B**) Weblogo of rarely-spliced junctions. (**C**) Degree of conservation of junction sequences pooled at discrete intervals of splicing efficiency. Each line represents a group of junction sequences pooled at discrete intervals of splicing efficiency, same as in (D). The entropy value (bits) at each position is calculated using Weblogo. (**D**) Positive correlation between CV and splicing efficiency. Each data point represents a group of junction sequences pooled at discrete intervals of splicing efficiency. The error bars represent standard error of the mean.

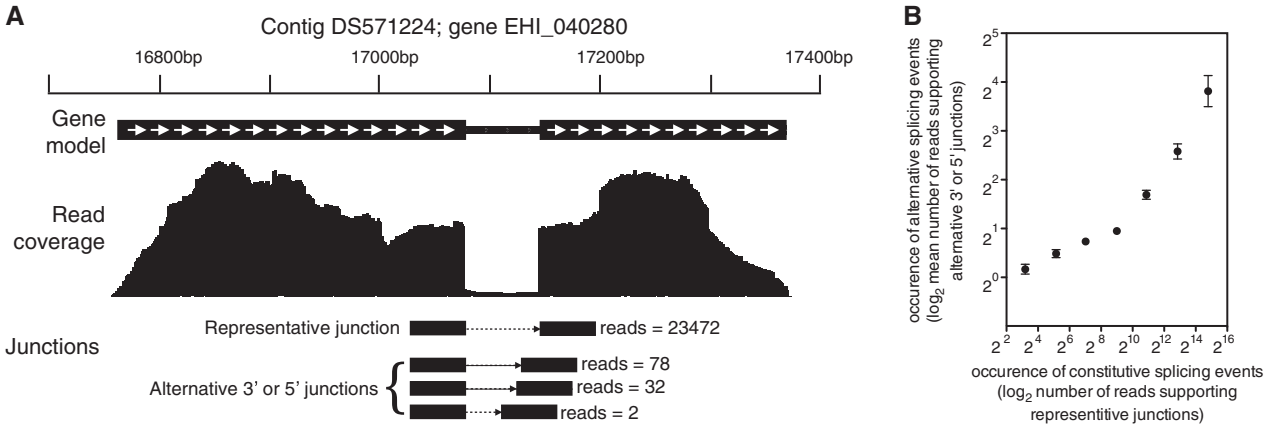

Figure 4. Alternative 5′ or 3′ splicing is more likely to occur in more frequently spliced junction clusters. (**A**) Example of a constitutive junction cluster. Representative and alternative 5′ or 3′ junctions within a constitutive junction cluster. Numbers of supported reads for each junction are indicated. (**B**) Positive correlation between numbers of reads supporting representative and alternative junctions within clusters. Each data point represents a group of junction clusters pooled at discrete intervals of number of reads supporting their representative junction. The error bars represent standard error of the mean.

Figure S1). It should be noted that alternative junctions mapped to non-unique regions were discarded in all downstream analyses to avoid potential artifacts. Next, we investigated the properties of these three types of alternative splicing events.

Alternative 5′ or 3′ splicing events are more likely to occur in more often spliced junctions

About 48% of the constitutive clusters contain more than one junction, suggesting alternative 5′ or 3′ splicing events are frequent. We reasoned if the majority of these alternative 5′ or 3′ splicing events were the stochastic noise of the constitutive splicing events, we should be able to observe a positive correlation between the occurrence of constitutive splicing event and alternative 5′ or 3′ splicing events. Occurrence of the constitutive splicing event within a cluster can be measured by number of reads supporting their representative junction, whereas occurrence of the alternative 5′ or 3′ splicing events can be measured by the total number of reads supporting these alternative junctions within the cluster. To this end, we grouped the junction clusters by discrete intervals of number of reads supporting their representative junction and plotted the mean number of reads supporting alternative junctions within these intervals (Figure 4B). We found the occurrence of constitutive splicing event is significantly correlated with the occurrence of alternative 5′ or 3′ splicing events (Figure 4B; Pearson's correlation coefficient = 0.96, $P < 0.0001$, 95% CI 0.79–0.99). These results imply the more often a constitutive junction is spliced, the higher probability for this junction to have alternative 5′ or 3′ splicing sites. If the majority of these alternative 5′ or 3′ splicing events were regulated physiologically, the overall occurrence of alternative 5′ or 3′ splicing events should be generally independent with the occurrence of constitutive splicing event, which is not the case. Therefore, most of these rarely-spliced alternative 5′ or 3′ splicing sites are likely to originate from the stochastic noise of the constitutive splicing events, rather than physiologically regulated.

Alternative exon skipping and intron creation are more likely to occur in more abundant mRNAs

Then, we investigated the properties of alternative exon skipping and intron creation events. Information of all alternative exon skipping and intron creation isoforms was summarized in Supplementary Table S2. About 7% ($n = 404$) of all gene models contain multiple constitutive introns, and ~20% ($n = 80$) of them showed evidence of exon skipping. All of these alternative exon skipping junctions are rarely spliced, with a median splicing efficiency of 0.0013, 5th and 95th percentile of 0.00011 and 0.0064%. We reasoned if most of these alternative exon skipping events are stochastic noise of the constitutive splicing events, more abundant mRNA (i.e. more frequent constitutive splicing events) should have higher probability of exon skipping. To this end, we compared the abundance of multiple intron mRNAs with ($n = 80$) and without ($n = 324$) exon skipping. We found the mRNAs with exon skipping are significantly more abundant than mRNAs without exon skipping (Figure 5A, $P < 0.0001$ in Student's *t*-test), suggesting alternative exon skipping events are more likely to occur in more abundant mRNAs.

On the other hand, ~17% ($n = 1237$) of the gene models were found to have alternative intron creation. To investigate the correlation between mRNA abundance and occurrence of intron creation events, we grouped all genes by discrete intervals of mRNA abundance and plotted the mean number of reads supporting alternative intron creation junctions in mRNAs within these intervals (Figure 5B). We found a significant positive correlation between mRNA abundance and occurrence of intron creation events (Figure 5B; Pearson's correlation coefficient = 0.99, $P < 0.0001$, 95% CI 0.95–0.99), suggesting these rare alternative intron creation events are more likely to occur in more abundant transcripts.

Estimation of the splicing error rate and identification of non-stochastic alternative splicing events

Based on the aforementioned observations, we conclude that majority of the alternatively splicing events are

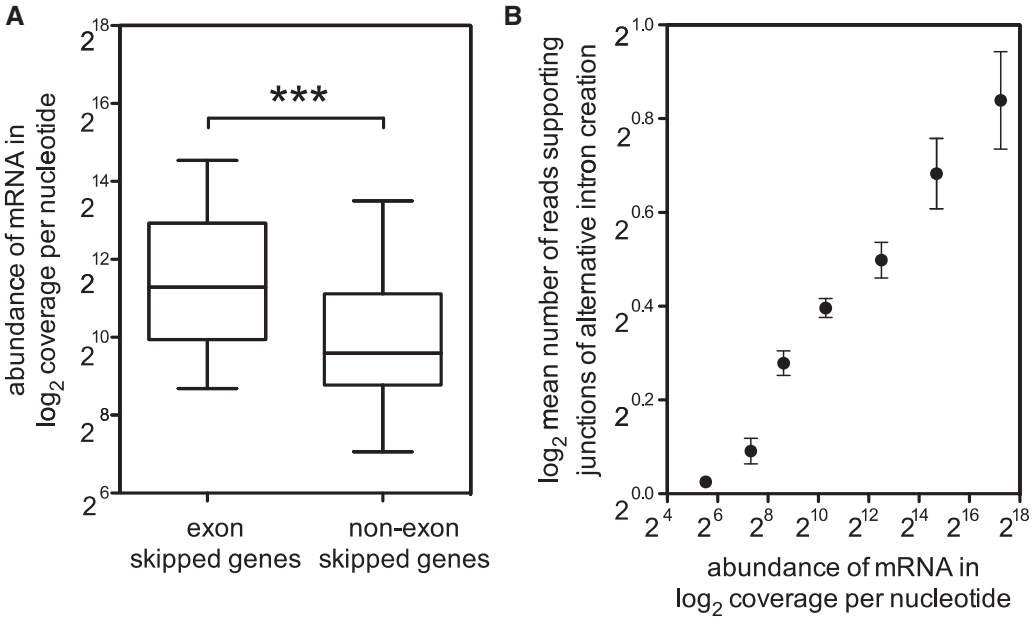

Figure 5. Alternative exon skipping and intron creation are more likely to occur in more abundant transcripts. (**A**) Abundance of multi-intron mRNA with and without exon skipping events. Asterisks: $P < 0.0001$ in Student's *t*-test. The 'boxes and whiskers' represent the 5th, 25th, 50th, 75th and 95th percentiles. (**B**) Positive correlation of mRNA abundance with the occurrence with intron creation events. Each data point represents a group of mRNA pooled at discrete intervals of mRNA abundance. The error bars represent standard error of the mean.

generally rare, and their occurrence is correlated with splicing site sequence, occurrence of constitutive splicing events and mRNA abundance. These results implied majority of these alternative splicing events seem to be derived from stochastic errors of splicing machineries in a probabilistic manner. We thus view functionally relevant alternative splicing events as physiologically regulated selection of spliced sites from a large pool of stochastically spliced alternative sites. Understanding the properties of these stochastic erroneous events certainly help us to identify the physiologically regulated events, which are more likely to be functionally relevant. To this end, we attempted to estimate the error rate for different types of alternative splicing events. Our goal is to identify the alternative splicing events that are substantially deviated from the estimated error rates, which are more likely to represent the non-stochastic alternative splicing events. First, we presume the alternative junctions with less conserved splicing site sequence, i.e. unconserved alternative junctions with low CV, are more likely to be solely stochastic. Then, we used the 5th percentile CV of all constitutive junctions (i.e. 0.7) as the cutoff to define these unconserved alternative junctions. These unconserved alternative junctions will be used for error rate estimations as described in the following section.

To estimate the error rate for alternative 5′ or 3′ splicing events, we first identified the constitutive junction clusters that contain only the aforementioned unconserved alternative junctions. Then, we calculated the ratio of read number supporting the alternative junctions to that of the constitutive junction in each cluster ($n = 128$ clusters). Theoretically, this ratio represents the number of unconserved alternative splicing events per constitutive splicing event (i.e. error rate). The observed median is 0.4%, with 5th and 95th percentile of 0.002% and 13%. Using the 95th percentile as the cutoff, we therefore defined non-stochastic alternative 5′ or 3′ splicing events as alternative junctions with supporting read number >13% of that of the constitutive junction ($n = 55$ out of 1647). We used the same cutoff to identify non-stochastic exon skipping events, i.e. alternative exon skipping junctions with supporting read number >13% of that of the skipped representative junctions ($n = 1$ out of 99). To estimate the error rate for intron creation events, we first identified all intron creation events that consist of unconserved alternative junctions ($n = 980$). Theoretically, the splicing efficiency of these alternative intron creation junctions represents the probability of a pre-mRNA molecule with an unconserved alternative intron created (i.e. error rate). The observed median is 0.1%, with 5th and 95th percentile of 0.0004% and 2%. Using the 95th percentile as the cutoff, we therefore defined non-stochastic intron creation events as alternative junctions with splicing efficiency >2% ($n = 136$ of 1667). It should be noted that these error rate estimations do not represent the inherent error rate of the splicing machineries themselves, but rather represent the observed error rate, resulting from the combinatory effects splicing machineries errors and RNA quality control machineries e.g. non-sense mediated decay pathway.

Possible origins of the non-stochastic alternative splicing events

To investigate whether these non-stochastic alternative splicing events tend to co-occur in multiple introns on the same transcript, we summarized the occurrence of non-stochastic alternative splicing sites in genes with

multiple constitutive introns (i.e. multi-intron genes). Non-stochastic alternative splicing sites were found in 31 of 447 multi-intron genes. Almost all ($n = 30$) of these 31 multi-intron genes have non-stochastic alternative splicing sites in only one of its constitutive introns. This observation suggests efficient alternative splicing sites do not tend to co-occur in multiple introns of the same transcript, implying a closer physical proximity between constitutive introns and the machineries needed for efficient alternative splicing does not seem to increase the probability of an intron being alternatively spliced.

To determine whether particular regulatory sequences were enriched around the splicing sites of these non-stochastic alternative intron creation events, we performed a discriminative discovery of motifs in flanking sequence of splicing site of non-stochastic alternative intron creation events (see 'Materials and Methods' section). As a control, we first used the flanking sequences of the constitutive splicing sites as the positive set and random coding sequences of non-intronic genes as the negative set. As expected, we identified the GUUG motif at the 5′ constitutive splicing sites, corresponding to the G UUGUU motif as described in Figure 3A. Next, we used the sequences flanking the splicing sites of the non-stochastic and stochastic alternative intron creation events as the positive and negative set, respectively. Again, we only identified the GUUG motif at the 5′ splicing site, and no additional significant motif was identified (data not shown). In fact, this observation reinforced our earlier observation that the splicing efficiency of a junction is correlated with the conservation of this GUUGUU motif at the 5′ splicing site sequence (Figure 3C and D). Therefore, we conclude that no additional regulatory sequence was found to be associated with non-stochastic alternative junctions except the GU UGUU motif at the 5′ splicing site.

Implications on the prevalence of functionally relevant alternative splicing isoforms

We consider the aforementioned 'non-stochastic' alternative splicing events ($n = 194$) are less likely to be solely stochastic and more likely to be physiologically regulated. These alternative splicing events and the supporting evidences in the six libraries were listed in Supplementary Table S3). To investigate the functional themes of these non-stochastic alternatively spliced genes, we summarized their gene ontologies and performed an enrichment test using GO::TermFinder. Although we did not find any significantly enriched functional themes, a summary of the gene ontologies of these non-stochastic alternatively spliced genes can be found in Supplementary Figure S6. In fact, >80% of these non-stochastic alternatively junctions ($n = 159$) are expected to cause frame shifts to the original ORF of the mRNA (Supplementary Table S3). Alternatively speaking, only 33 of them are expected to generate alternative transcripts without premature stop codons. If we presume coding potential (i.e. with or without premature stop codons) of an alternative transcript is crucial to its functional relevance, then the prevalence of functionally relevant alternative splicing isoforms in *E. histolytica* would be limited. To summarize, despite the pervasiveness of alternative splicing events, most of them are likely to be stochastic, and the functional impact of alternative splicing in *E. histolytica* is minimal.

Majority of the identified poly(A) sites are genuine

Now, we shift our focus to alternative polyadenylation. Here, we confine our discussions to poly(A) site of mRNAs only. To assess whether the identified poly(A) sites are genuine, we plot the total occurrence of poly(A) sites around the stop codons of all gene models (Figure 6A). As expected, most of the poly(A) sites fall within 100 nt downstream. We also cross-validated our results with the annotated ends of 3′UTR in AmoebaDB by searching for the closest poly(A) sites within ±200 nt of these annotated 3′UTR ends. More than 90% (584 of 645) of these annotated ends of 3′UTR overlapped with the identified poly(A) sites within ±10 nt, and 399 of them were confirmed to the exact nucleotide (Figure 6B). These results suggest most of the identified poly(A) sites are genuine. Totally, we identified 331 306 poly(A) reads associated with mRNAs, allowing us to define the poly(A) sites of ~62% of gene models ($n = 4509$). We noticed that 62% of mRNA poly(A) sites coverage is relatively low, which is likely because of the underrepresentation of RNA fragments from transcript ends in the libraries. Libraries enriched in poly(A) RNA fragments, e.g. PolyA-Seq (8), would certainly increase the coverage of poly(A) sites. A gene-by-gene summary of poly(A) sites at 3′UTR can be found in Supplementary Table S5. The median distance from the poly(A) sites to the corresponding stop codon is 21 nt, with 5th and 95th percentile of 11 and 75 nt.

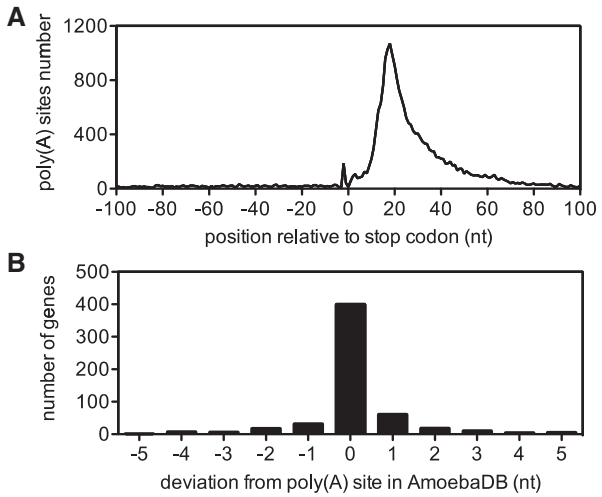

Figure 6. Majority of the identified poly(A) sites are genuine. (**A**) Occurrence of poly(A) sites around the stop codons of all gene models. Most of the poly(A) sites fall within 100 nt downstream of stop codon. (**B**) Overlapping between the 3′UTR ends annotated in AmoebaDB and poly(A) sites identified in our data set. A total of 584 of 645 of AmoebaDB annotated 3′UTR ends overlapped with the identified poly(A) sites within ±10 nt.

Microheterogeneity in poly(A) cleavage sites are stochastic

Microheterogeneity in poly(A) cleavage sites in higher eukaryotes is well documented (6). Here, we quantified such microheterogeneity observed in *E. histolytica*. First, we grouped the poly(A) sites into clusters by allowing certain cutoffs on the maximum distance between sites within a cluster. About 85% of sites can be grouped as clusters if we allow a cutoff of maximum 12 nt between sites, and further increasing the cutoff did not increase the percentage of clustered site significantly (Figure 7A). The result suggests most of the sites were in proximity to other sites within a range of 12 nt, and therefore, we chose 12 nt as the cutoff to define poly(A) clusters. In fact, 95% of poly(A) clusters are sized <20 nt (Figure 7B). Information of all poly(A) site clusters was summarized in Supplementary Table S4. Each poly(A) cluster was

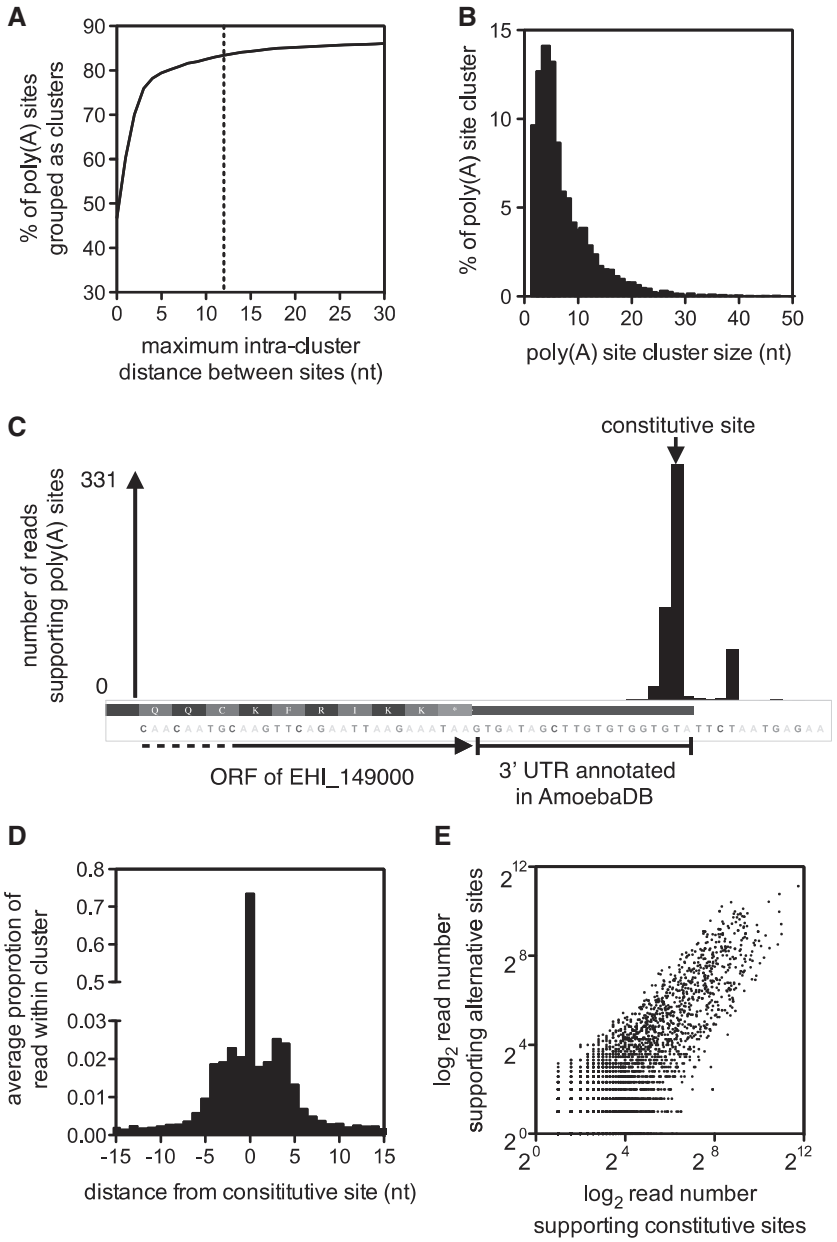

Figure 7. Microheterogeneity in poly(A) cleavage sites are stochastic. (**A**) Defining microheterogeneity [i.e. poly(A) site clusters] using maximum intra-cluster distance. Further increasing the maximum intra-cluster distance >12 nt (dotted line) did not substantially increase the percentage of clustered poly(A) sites. (**B**) Defining long-range heterogeneity based on poly(A) site cluster size. About 98% of clusters are <25 nt; therefore, a cutoff of 25 nt was chosen to define long-range heterogeneity. (**C**) Example of microheterogeneity within a poly(A) site cluster. The constitutive site was indicated with an arrow, while the other sites are alternative sites. (**D**) Histogram of the proportion of reads supporting constitutive and alternative cleavage sites within poly(A) site clusters. Position 0 corresponds to constitutive cleavage site. The average proportion of reads supporting constitutive events in all clusters is >0.75. (**E**) Positive correlation between the number of reads supporting constitutive and alternative cleavage site within clusters. Each data point represents a poly(A) site cluster.

represented by a constitutive site (i.e. the 'peak' with most number of reads), and the rest were termed alternative sites (Figure 7C). We defined microheterogeneity as the alternative sites within clusters. To investigate the usage of constitutive site versus alternative sites within clusters, we calculated the proportion of reads supporting the constitutive site and its surrounding sites in each cluster. Then, we plotted the average proportion of reads supporting each relative positions of all clusters (Figure 7D). Most of the alternative sites were located within ±5 nt to the corresponding constitutive site, reflecting the inherent limits of the range of such microheterogeneity (Figure 7D). These alternative sites contributed less than a proportion of 0.25 of all cleavage events within a cluster on average, as the average proportion of reads supporting constitutive sites is >0.75 (Figure 7D). These results suggest that although microheterogeneity of poly(A) cleavage site seem to be ubiquitous, majority of the cleavage events (>75%) still occur on the constitutive site. To this end, we reasoned if such microheterogeneity is because of stochastic noise of the constitutive cleavage events, we should be able to observe a positive correlation between the occurrence of constitutive cleavage events and the amount of microheterogeneity within clusters. We measured these two properties by the number of reads supporting the constitutive sites and the total number of reads supporting the alternative sites within a cluster, respectively. As shown in Figure 7E, these two parameters are highly correlated (Pearson's correlation coefficient = 0.81, $P < 0.0001$, 95% CI 0.80–0.82), implying the extent of microheterogeneity within a cluster increases with the occurrence of constitutive cleavage events. Therefore, such microheterogeneity is likely to be because of stochastic noise of the constitutive cleavage events within clusters.

Long-range alternative poly(A) events in mRNA are limited

Now, we consider long-range heterogeneity of poly(A) events, including multiple distinct poly(A) clusters in 3′UTR and poly(A) clusters within ORF (see 'Materials and Methods' section). Multiple distinct poly(A) clusters in 3′UTR were observed in 51 genes, and poly(A) clusters within ORF were observed in 59 genes, while three genes have both types of alternative poly(A) events (i.e. totally 107 genes were affected, listed in Supplementary Tables S5 and S6). Gene ontologies of these alternatively polyadenylated genes were summarized in Supplementary Figure S7. A list of these alternatively polyadenylated genes can be found in Supplementary Table S5. Further analyses of sequence properties of these long-range alternative poly(A) clusters (discussed later) suggest these alternative poly(A) events are mostly genuine. It is noted that 20 of the 59 genes with poly(A) clusters mapped within ORF have no poly(A) cluster mapped within 3′ UTR. To estimate the extent alternative polyadenylation in *E. histolytica*, we calculated the percentage of genes with long-range alternative poly(A) events in two pools of genes, (i) genes with polyadenylation sites mapped within 3′UTR ($n = 87$ out of 4509 genes, ~1.9%); and (ii) genes with at least 10 supporting reads in its prominent poly(A) cluster within 3′UTR (i.e. please refer to 'mostRdPolyARdNum' column in Supplementary Table S5; $n = 51$ of 1995 genes, ~2.4%). These results suggest that although long-range heterogeneity of poly(A) events is observed in *E. histolytica*, these alternative poly(A) events seems to only occur in a limited proportion of genes (~1.9 to ~2.4%). In fact, given its relatively short 3′UTR (median of 21 nt), we do not expect too much long-range alternative polyadenylation in 3′UTR of mRNAs in *E. histolytica*, in contrast to the pervasive alternative polyadenylation in other characterized model organisms (2).

Sequence characteristics surrounding polyadenylation sites

Having summarized the microheterogeneity and long-range heterogeneity of poly(A) sites, we next explore the base compositions and sequence motifs surrounding poly(A) sites of coding genes. We aligned 200-nt sequence surrounding all poly(A) clusters in the 3′ UTR of mRNAs ($n = 5018$). The base composition profile is characterized by a narrow A-rich peak at around −20 nt, and a broad U-rich region surrounding the cleavage site (Figure 8A). The A-rich peak at −20 nt corresponds to the location of canonical polyadenylation signal sequence found in other model organisms (8). It should also be noted that the enrichment of C at −1 nt (Figure 8A) supports the observation that a CA dinucleotide immediately 5′ to the cleavage site is preferred but not absolutely required (38). Using DREME, we identified an AAWUDA motif at around the −20 nt (*E*-value = 4.4^{-368}) (Figure 8B and C), resembling the canonical polyadenylation signal AAUAAA in mammalian species (8). This AAWUDA motif (Figure 8C) is highly position-specific at −20 nt (Figure 8B). In addition, we also identified U-rich motifs at immediate upstream (UUUUUW, *E*-value = 4.8^{-74}) and downstream (UUUUWW, *E*-value = 2.8^{-328}) of the poly(A) site (data not shown). These two U-rich motifs correspond to the broad U-rich region in the base composition profile surrounding the poly(A) site (Figure 8A). In summary, three sequence characteristics, including AAWUDA motif at −20 nt, enrichment of C at −1 nt and a broad U-rich region surrounding the poly(A) site, seem to play roles in determining the cleavage site of polyadenylation.

Occurrence of AAWUDA motif in different types of poly(A) clusters

To this end, we investigated the occurrence of AAWUDA motif and the proportion of its variants in different types of poly(A) clusters. We categorized the poly(A) clusters into three types, (i) single poly(A) clusters at 3′UTR ($n = 4173$); (ii) multiple poly(A) clusters at 3′UTR ($n = 102$); and (iii) poly(A) clusters within ORF ($n = 59$). Sequences of 15-nt region at around 20 nt upstream of these poly(A) clusters were extracted, and the occurrence of AAWUDA motif was counted. As a negative control, random genomic sequences of the same length (15 nt, $n = 10\,000$) were also counted. Permutated sequences from each type were also counted as

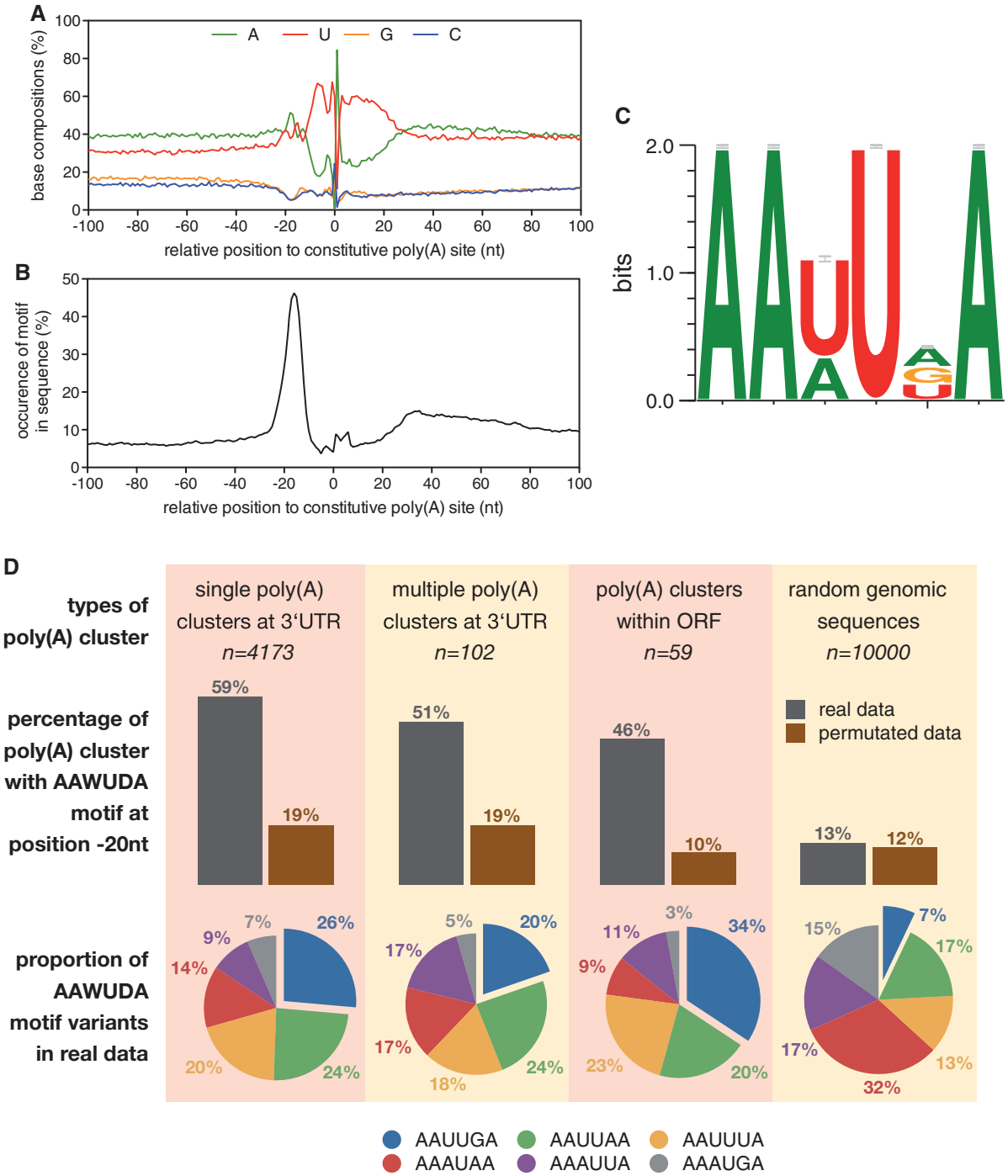

Figure 8. Sequence characteristics surrounding polyadenylation sites. (**A**) Base compositions surrounding poly(A) sites of coding genes. (**B**) Positional occurrence of the AAWUDA motif along the sequences surrounding poly(A) sites. This motif is enriched around position −20 nt. Position 0 corresponds to constitutive cleavage site. (**C**) Weblogo of the AAWUDA motif identified in DREME. The entropy value (bits) at each position is calculated using Weblogo. (**D**) Occurrence of the AAWUDA motif between position −25 nt to −10 nt of different types of poly(A) cluster. Middle panel: percentage presence of AAWUDA motif in real and permutated sequence of the corresponding poly(A) cluster types. Lower panel: proportion of AAWUDA motif variants in real sequence of the corresponding poly(A) cluster types. Random genomic sequences of same length were used as control.

background controls (see 'Materials and Methods' section). About 59% of the real sequences extracted from single poly(A) clusters at 3′UTR were found to contain AAWUDA motif, which is substantially higher than both that of the corresponding permutated sequences and the random genomic sequences (Figure 8D), suggesting the presence of AAWUDA motif in the upstream of these poly(A) clusters is non-random. This result also implies the AAWUDA motif is absent in ∼41% of these poly(A) clusters. It should be noted that we only

considered the six predefined variants within a narrow window of 15 nt, and therefore, this percentage of 'signaless' poly(A) clusters could be over-estimated. Further analyses of these 'signaless' poly(A) clusters did not yield any significant sequence patterns. Nonetheless, it was reported that a canonical polyadenylation signal sequence might not be an absolute prerequisite for a polyadenylation event, as ~13% of mRNA isoform poly(A) sites in *Caenorhabditis elegans* (9) do not contain any recognizable polyadenylation signal sequences. About 51 and 46% of multiple poly(A) clusters at 3′UTR and poly(A) clusters within ORF were found to contain the AAWUDA motif, respectively, which are comparable with that of the single poly(A) clusters at 3′UTR and substantially higher than that of the corresponding permutated sequences, suggesting these long-range alternative poly(A) clusters are likely to be genuine. Finally, the proportion of variant AAUUGA (blue slices in pie charts of Figure 8D) in all types of poly(A) clusters (~20–34%) are consistently and substantially higher than that of the random genomic sequences (~7%), further supporting the occurrence of AAWUDA motif around different types of poly(A) clusters are non-random. Taken together, these data further support that the AAWUDA motif is a genuine polyadenylation signal sequence in *E. histolytica*.

Variability of alternative polyadenylation and splicing between HM1:IMSS and Rahman strain

To investigate the variability of alternative poly(A) clusters between the two strains, we calculated the change of proportion of these poly(A) events in the total number of poly(A) events of the corresponding genes between the two strains. Under our criteria (see 'Materials and Methods' section), only 10 of the 107 alternatively polyadenylated genes showed evidence of substantial shifts in proportion of alternative poly(A) events (≥2-fold). Figure 9 showed an example of substantial increase of poly(A) event within ORF of gene EHI_161070 in Rahman strain. In summary, although we observed variability in alternative poly(A) events between the two strains, such variability is limited to only a few genes [i.e. shift in poly(A) isoform proportion, $n = 10$]. Details of the inter-strain and inter-library variability of alternative poly(A) clusters were listed in Supplementary Table S7.

Next, to investigate the variability of the non-stochastic alternative splicing isoforms between the two strains, we compared the number of reads supporting these alternative junctions ($n = 194$) among the six libraries. Here, we define strain-specific alternative junctions as junctions with supporting reads in one or more replicates of a strain but without supporting reads in all replicates of the other strain. We found ~85% ($n = 163$ of 194) of these alternative junctions were present in both strains, and thus, ~15% of these alternative junctions ($n = 31$ of 194) seem to be strain specific. However, 26 of these 31 strain-specific alternative junctions have <10 supporting reads (in sum of three replicates), and therefore, the strain specificity of these 26 alternative junctions is still inconclusive. Alternatively speaking, only five alternative junctions can be confidently classified as strain specific. To investigate whether the alternative isoforms are differentially expressed, we tested for differential expression between the two strains using Cufflinks (see 'Materials and Methods' section). Only one alternative isoform showed statistically significant (false discovery rate ≤5%) differential expression.

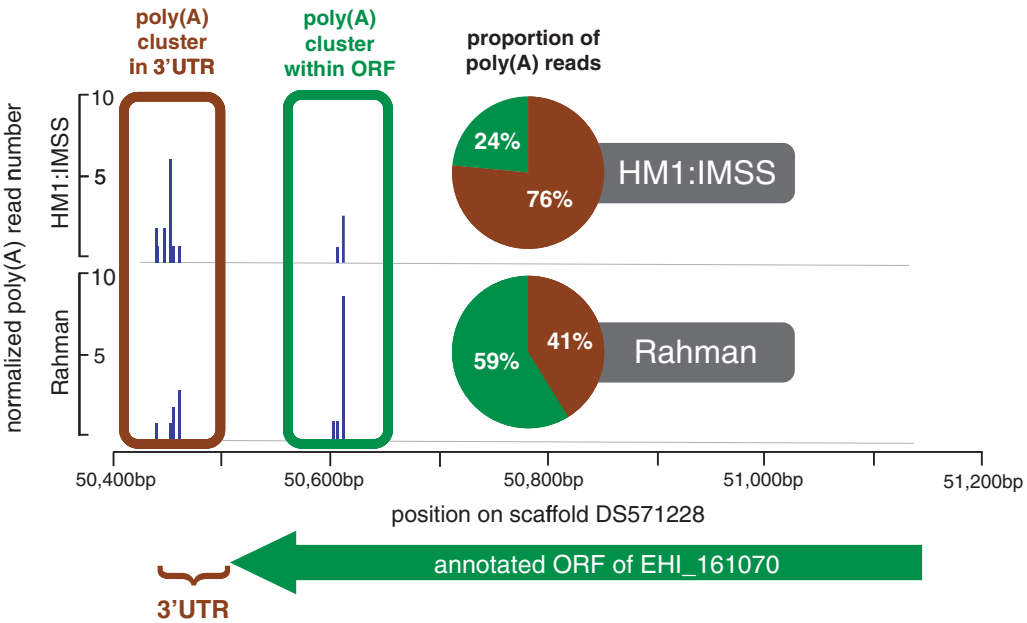

Figure 9. Shift in proportion of alternative poly(A) events between HM1:IMSS and Rahman strain. Normalized poly(A) read number supporting a poly(A) cluster in 3′UTR (brown box) and an alternative poly(A) cluster within ORF (green box) of gene EHI_161707 in HM1:IMSS (upper track) and Rahman strain (lower track) was plotted. Proportion of poly(A) reads supporting the two clusters was plotted as pie charts and the color correspond to the boxes on the left. The proportion of reads in alternative poly(A) cluster in Rahman is >2-fold higher than in HM1:IMSS strain.

We also empirically compared the fold change of number of reads supporting the alternative junctions between the two strains (see 'Materials and Methods' section). Only 21 alternative junctions showed fold change of supporting reads ≥3 between the two strains. Details of the inter-strain variability of alternative splicing isoforms were listed in Supplementary Table S8. In summary, although we observed differences in alternative splicing between the two strains, such differences are limited to a small number of genes (i.e. strain-specific splicing isoform, $n = 5$; differentially expressed splicing isoforms, $n = 21$).

To this end, we summarized the GO terms of the genes that exhibit variability of alternative polyadenylation or splicing between the two strains ($n = 10 + 5 + 21$ as described earlier). Although we did not find any significantly enriched functional themes, a summary of the gene ontologies of these genes can be found in Supplementary Figure S8. In conclusion, inter-strain variability in alternative splicing and polyadenylation was only observed in a small fraction of genes, and thus, we expect the contribution of alternatively splicing and polyadenylation to the phenotypic differences (e.g. virulence) between the two strains is likely to be limited. Experimental validation of the previously listed genes might shed light on this aspect.

Identification of novel coding transcripts

The genome of *E. histolytica* was sequenced in 2005 (19), and was then re-assembled and re-annotated recently (21). The current assembly, which consists of 1496 scaffolds, is often considered as 'unfinished', i.e. it might contain numerous misassembled regions and wrongly annotated gene models (20). Therefore, as a part of the effort to improve the gene model annotations, we identified 181 novel coding transcripts. We considered these novel coding transcripts *bona fide*, as they were curated manually to ensure the start and stop codons are located at the extremities of the transcripts, based on the fact that UTR of *Entamoeba* mRNA is extremely short (39). Most of these novel ORFs (i.e. 159 of 181) are shorter than 300 nt (average ≈268 nt), reflecting the fact that ORFs of <100 codons were basically ignored in previous gene predictions (21). Identities of the predicted proteins were then annotated using InterProScan and BLAST. Interestingly, ~30 and ~16% of these short peptides are predicted to contain transmembrane domains and coiled-coil domains, respectively. In addition, a number of small peptides with well-known functions were also identified, e.g. G-protein subunit gamma (EHI_C00092), amoebapore (EHI_C00062), ferredoxin (EHI_C00161), thioredoxin (EHI_C00011), etc. In particular, amoebapores, which are ~77-residue amphipathic peptides that are able to form transmembrane pores with a hydrophilic hole in membranes of target cells, are thought to be an important virulence factor (40). A better annotated catalog of these small peptides might therefore lead to the discovery of novel virulence factors. All of these novel coding transcripts (except EHI_C00151) are expressed in both HM1:IMSS and Rahman strains. Functional annotations, as well as the read counts of these genes in both strains, were listed in Supplementary Table S9.

Validation and revision of existing gene models

Finally, we revised 720 existing gene models and validated most of the rest, defining a set of 7312 *bona fide* gene models (including the 181 novel genes mentioned previously). Alternatively speaking, ~15% ($n = 1220$) of the existing gene models could not be validated, while ~73% ($n = 892$) of these 'invalid' models were because of conflicts with inherent criterion (see 'Materials and Methods' section), i.e. either lack of a complete ORF (e.g. pseudogene) or located at the proximity of ambiguous genomic regions (e.g. scaffold breaks). In fact, majority of these unvalidated gene models are repetitive, with only <25% of them being unique (see 'Materials and Methods' section, Supplementary Figure S9). It thus agrees with fact that (i) repetitive sequences tend to create breaks in the scaffolds; and (ii) pseudogenes are often repetitive. About 44% of the modified gene models involve changes in junctions (i.e. novel, missed and modified junctions in Supplementary Figure S10A), while most of them ($n = 189$) involve incorporation of new junctions into the gene models (Supplementary Figure S10B). In addition, based on manual curation, 59 and 12 existing gene models were split (Supplementary Figure S10C) and fused (Supplementary Figure S10D), respectively. On the other hand, ~47 and 5% of the modified gene models involved extension (Supplementary Figure S10E) and shortening (Supplementary Figure S10F) of transcript ends, respectively. All finalized gene models, as well as the RNA-Seq data, can be found in the next release of AmoebaDB (25). Information on gene model revision was summarized in Supplementary Table S10. In summary, we defined a set *bona fide* gene models ($n = 7312$), and ~62% ($n = 4509$) of them are annotated with 3′UTR. These annotations represent a major leap forward in the genome biology of the parasite since its genome was sequenced, providing valuable resources to the research community.

DISCUSSION

Distinguishing physiologically regulated events from inherently stochastic events

Our primary aim was to assess the extent of alternative splicing and polyadenylation, as well as their functional relevance, in *E. histolytica*. Functional significances of alternative splicing and polyadenylation are undoubted, but only limited to a handful of well-studied examples (12,13). Although numerous studies demonstrated the pervasiveness of alternative splicing and polyadenylation in higher eukaryotes, little is known about the functional relevance of the resulting RNA isoforms. Here, we emphasize the importance of quantifying such heterogeneity to distinguish physiologically regulated events from inherently stochastic events, which both result in mRNA diversity. We therefore estimated the error rates of splicing and identified the alternative splicing events that are less likely to be solely stochastic. For instance, only 33 of

them are not likely to be solely stochastic and expected to generate alternative transcripts without premature stop codons, suggesting alternative splicing is not likely to dramatically expand the proteome diversity of *E. histolytica*. On the other hand, we quantified the extent of stochastic microheterogeneity of polyadenylation and defined a cutoff of 25 nt to identify long-range heterogeneity that is less likely to solely stochastic. In fact, long-range alternative polyadenylation seems to occur in small proportion of genes (estimated to be ~1.9 to ~2.4%), suggesting alternative polyadenylation only plays a limited role in expanding the transcriptome diversity of *E. histolytica*. Also, variability of the non-stochastic alternatively spliced and long-range alternative polyadenylated isoforms are limited between HM1:IMSS and Rahman strain. In summary, we expect the functional impacts of these alternative splicing and polyadenylation events are only limited to a small proportion of genes in this organism.

Evidence and consequence of noisy splicing

Most, if not all, biological processes, including splicing, have inherent error rates. To the best of our knowledge, the number of analyses of splicing noise based on empirical data is limited. Melamud and Moult (14) analyzed the human EST libraries and showed that the number of isoforms and their abundance can be predicted through a simple stochastic noise model that takes into account only the number of introns and expression level of a gene. Alternatively speaking, they found the number of detected alternative splicing reactions increases with the total number of observed splicing reaction, strongly supportting the hypothesis that most alternative splicing is a consequence of stochastic noise in the splicing machinery and has no functional significance. More recently, using RNA-Seq, Pickrell *et al*. demostrated the existence of a large class of low abundance alternative isoforms. They found little evidence of evolutionary conservation in the splicing sites of these isoform, and splicing error rate is correlated with intron length, suggesting that the majority are because of erroneous splicing site choice (16). In this study, there are three main observations supporting the notion that the majority of alternative junctions are consequence of splicing noise, including (i) splicing efficiency is correlated with splicing site sequence conservation (Figure 3); (ii) most of the rarely-spliced alternative 5′ or 3′ splicing sites are dependent on the occurrence of the constitutive splicing events (Figure 4); and (iii) exon skipping and intron creation are more likely to occur in more abundant transcripts (Figure 5). Although we expect most of these isoforms are likely to be non-functional, from a evolutionary standpoint, these isoforms might provide the substrate for a yet to be known selection process, which selects for a functionally relevant isoform and ultimately contributes to the emergence of novel biological properties. Skandalis *et al*. (41) investigated the alternative splicing patterns of DNA polymerase β among primate species and found the frequency of unproductive alternative splicing is correlated well with life history of parameters, such as the maximal longevity of each species. They speculate unproductive alternative splicing may have adaptive significance even if the transcripts themselves are non-functional. In any case, quantification of these unproductive alternative splicing would be the stepping stone to understand their evolutionary significance.

Heterogeneity of poly(A) site in an early branching eukaryote

Tian *et al*. (42) reported that heterogeneity of poly(A) cleavage usually occurs within 24 nt after the 5′-most cleavage site in human and mouse genes. Our observation is comparable with theirs, with ~97% of poly(A) clusters sized <24 nt. Tian *et al*. (42) also showed that the number of cleavage sites is correlated with the number of supporting cDNA/ESTs, suggesting such microheterogeneity is generally stochastic. Here, we further quantified this concept by demonstrating a strong correlation between the occurrence of constitutive and alternative cleavage events, suggesting the inherent stochastic errors of the more frequent constitutive cleavage events are the origin of the less frequent alternative cleavage events. These results suggest the origin of microheterogeneity in poly(A) cleavage site seems to be universal from lower to higher eukaryotes. However, the pervasiveness of long-range heterogeneity of poly(A) site in higher eukaryotes, e.g. 69.1% of human genes have multiple poly(A) clusters (8), seems to be lacking in *E. histolytica*. In fact, only a small proportion of genes (estimated to be ~1.9 to ~2.4%) were found to have long-range heterogeneity of poly(A) site in *E. histolytica*. The relative short 3′UTR (median 21 nt) and the lack of extensive alternative polyadenylation might imply the regulatory roles of 3′UTR in *E. histolytica* are relatively limited when compared with other higher eukaryotes (13).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–10 and Supplementary Figures 1–10.

ACKNOWLEDGEMENTS

A.B. thanks the Department of Biotechnology, Government of India, for support.

FUNDING

French National Research Agency [ANR-2010-GENM-011-01, GENAMIBE to N.G.]; French National Research Agency (to C.C.H., M.K. and M.D.); CSIR-SRF fellowship (to S.D.); BNP Paribas (CIB for Health and Medical Research) (to N.G. and A.B.). Funding for open access charge: French National Research Agency [ANR-2010-GENM-011-01, GENAMIBE to N.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Keren,H., Lev-Maor,G. and Ast,G. (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**, 345–355.
2. Ozsolak,F., Kapranov,P., Foissac,S., Kim,S.W., Fishilevich,E., Monaghan,A.P., John,B. and Milos,P.M. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, **143**, 1018–1029.
3. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
4. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
5. Filichkin,S.A., Priest,H.D., Givan,S.A., Shen,R., Bryant,D.W., Fox,S.E., Wong,W.K. and Mockler,T.C. (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res*, **20**, 45–58.
6. Pauws,E., van Kampen,A.H., van de Graaf,S.A., de Vijlder,J.J. and Ris-Stalpers,C. (2001) Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res.*, **29**, 1690–1694.
7. Iseli,C., Stevenson,B.J., de Souza,S.J., Samaia,H.B., Camargo,A.A., Buetow,K.H., Strausberg,R.L., Simpson,A.J., Bucher,P. and Jongeneel,C.V. (2002) Long-range heterogeneity at the 3' ends of human mRNAs. *Genome Res.*, **12**, 1068–1074.
8. Derti,A., Garrett-Engele,P., Macisaac,K.D., Stevens,R.C., Sriram,S., Chen,R., Rohl,C.A., Johnson,J.M. and Babak,T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, **22**, 1173–1183.
9. Mangone,M., Manoharan,A.P., Thierry-Mieg,D., Thierry-Mieg,J., Han,T., Mackowiak,S.D., Mis,E., Zegar,C., Gutwein,M.R., Khivansara,V. *et al.* (2010) The landscape of C. elegans 3'UTRs. *Science*, **329**, 432–435.
10. Wu,X., Liu,M., Downie,B., Liang,C., Ji,G., Li,Q.Q. and Hunt,A.G. (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl Acad. Sci. USA*, **108**, 12533–12538.
11. Shen,Y., Ji,G., Haas,B.J., Wu,X., Zheng,J., Reese,G.J. and Li,Q.Q. (2008) Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res.*, **36**, 3150–3161.
12. Kalsotra,A. and Cooper,T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, **12**, 715–729.
13. Gilmartin,G.M. (2005) Eukaryotic mRNA 3' processing: a common means to different ends. *Genes Dev.*, **19**, 2517–2521.
14. Melamud,E. and Moult,J. (2009) Stochastic noise in splicing machinery. *Nucleic Acids Res.*, **37**, 4873–4886.
15. Sorek,R., Shamir,R. and Ast,G. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**, 68–71.
16. Pickrell,J.K., Pai,A.A., Gilad,Y. and Pritchard,J.K. (2010) Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, **6**, e1001236.
17. Stanley,S.L. Jr (2003) Amoebiasis. *Lancet*, **361**, 1025–1034.
18. Song,J., Xu,Q., Olsen,R., Loomis,W.F., Shaulsky,G., Kuspa,A. and Sucgang,R. (2005) Comparing the Dictyostelium and Entamoeba genomes reveals an ancient split in the Conosa lineage. *PLoS Comput. Biol.*, **1**, e71.
19. Loftus,B., Anderson,I., Davies,R., Alsmark,U.C., Samuelson,J., Amedeo,P., Roncaglia,P., Berriman,M., Hirt,R.P., Mann,B.J. *et al.* (2005) The genome of the protist parasite Entamoeba histolytica. *Nature*, **433**, 865–868.
20. Loftus,B.J. and Hall,N. (2005) Entamoeba: still more to be learned from the genome. *Trends Parasitol.*, **21**, 453.
21. Lorenzi,H.A., Puiu,D., Miller,J.R., Brinkac,L.M., Amedeo,P., Hall,N. and Caler,E.V. (2010) New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. *PLoS Negl. Trop. Dis.*, **4**, e716.
22. Zamorano,A., Lopez-Camarillo,C., Orozco,E., Weber,C., Guillen,N. and Marchat,L.A. (2008) 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.
23. Wilihoeft,U., Campos-Gongora,E., Touzni,S., Bruchhaus,I. and Tannich,E. (2001) Introns of Entamoeba histolytica and Entamoeba dispar. *Protist*, **152**, 149–156.
24. Davis,P.H., Schulze,J. and Stanley,S.L. Jr (2007) 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.
25. Aurrecoechea,C., Barreto,A., Brestelli,J., Brunk,B.P., Caler,E.V., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G. *et al.* (2011) AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res.*, **39**, D612–D619.
26. Lorenzi,H., Thiagarajan,M., Haas,B., Wortman,J., Hall,N. and Caler,E. (2008) Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species. *BMC Genomics*, **9**, 595.
27. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
28. Dimon,M.T., Sorber,K. and DeRisi,J.L. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One*, **5**, e13875.
29. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
30. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
31. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
32. Lee,J.Y., Park,J.Y. and Tian,B. (2008) Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. *Methods Mol. Biol.*, **419**, 23–37.
33. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
34. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
35. Le,S., Josse,J. and Husson,F. (2008) FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.*, **25**, 1–18.
36. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
37. Roy,S.W. and Penny,D. (2007) Intron length distributions and gene prediction. *Nucleic Acids Res.*, **35**, 4737–4742.
38. Chen,F., MacDonald,C.C. and Wilusz,J. (1995) Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res.*, **23**, 2614–2620.
39. Bhattacharya,A., Satish,S., Bagchi,A. and Bhattacharya,S. (2000) The genome of Entamoeba histolytica. *Int. J. Parasitol.*, **30**, 401–410.
40. Bracha,R., Nuchamowitz,Y. and Mirelman,D. (2002) Amoebapore is an important virulence factor of Entamoeba histolytica. *J. Biosci.*, **27**, 579–587.
41. Skandalis,A., Frampton,M., Seger,J. and Richards,M.H. (2010) The adaptive significance of unproductive alternative splicing in primates. *RNA*, **16**, 2014–2022.
42. Tian,B., Hu,J., Zhang,H. and Lutz,C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, **33**, 201–212.