Messenger RNAs with large numbers of upstream open reading frames are translated via leaky scanning and reinitiation in the asexual stages of *Plasmodium falciparum*

Chhaminder Kaur, Mayank Kumar* and Swati Patankar

Department of Biosciences & Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

Abstract

The genome of *Plasmodium falciparum* has one of the most skewed base-pair compositions of any eukaryote, with an AT content of 80–90%. As start and stop codons are AT-rich, the probability of finding upstream open reading frames (uORFs) in messenger RNAs (mRNAs) is high and parasite mRNAs have an average of 11 uORFs in their leader sequences. Similar to other eukaryotes, uORFs repress the translation of the downstream open reading frame (dORF) in *P. falciparum*, yet the parasite translation machinery is able to bypass these uORFs and reach the dORF to initiate translation. This can happen by leaky scanning and/or reinitiation.

In this report, we assessed leaky scanning and reinitiation by studying the effect of uORFs on the translation of a dORF, in this case, the luciferase reporter gene, and showed that both mechanisms are employed in the asexual blood stages of *P. falciparum*. Furthermore, in addition to the codon usage of the uORF, translation of the dORF is governed by the Kozak sequence and length of the uORF, and inter-cistronic distance between the uORF and dORF. Based on these features whole-genome data was analysed to uncover classes of genes that might be regulated by uORFs. This study indicates that leaky scanning and reinitiation appear to be widespread in asexual stages of *P. falciparum*, which may require modifications of existing factors that are involved in translation initiation in addition to novel, parasite-specific proteins.

Introduction

Malaria affects millions of people in tropical and sub-tropical regions of the world. Over the years, many attempts have been made to control the disease and recently, the World Health Organization renewed the call towards the global eradication of malaria by 2030 (World Malaria Report, 2018). These efforts could be thwarted through resistance to anti-malarial drugs and as a result, new drug targets are still being identified, especially for the most virulent species of the parasite that causes malaria: *Plasmodium falciparum*. In this parasite, an essential pathway proposed as a drug target is the protein synthesis machinery (Goodman et al., 2016). *Plasmodium falciparum* employs eukaryotic protein synthesis machinery for translation of nuclear-encoded mRNAs, however, an endosymbiotic organelle, the apicoplast, carries out trans-translation using machinery that resembles prokaryotes (Fichera and Roos, 1997; Roy et al., 1999; Chaubey et al., 2005) and is the target of several anti-malarial drugs (Goodman et al., 2016; Pasaje et al., 2016). Interestingly, the cytosolic translation machinery of *P. falciparum* also exhibits differences from that of the human host (Jackson et al., 2011; Wong et al., 2014) and hence, has also been proposed to be a drug target (Wong et al., 2017; Sheridan et al., 2018). Another noteworthy distinction between the cytosolic translation of the parasite and its host is a result of the high overall content of adenine and thymine (80.6%) of the *P. falciparum* genome that rises to 90% in the intergenic regions (Gardner et al., 2002). As start and stop codons are AT-rich, this biased sequence leads to the presence of numerous upstream open reading frames (uORFs) in the 5′ leader sequences of messenger RNAs (mRNAs) (Caro et al., 2014; Kumar et al., 2015).

Unlike human and mouse genomes, from which ~50% of the transcribed mRNAs would be expected to contain one or more uORFs (Calvo et al., 2009; Ye et al., 2015), 96.5% of *P. falciparum* mRNAs are predicted to contain at least one uORF per coding sequence (CDS) (Srinivas et al., 2016). An average of ~4 uORFs has been predicted within 350 nucleotides upstream of the start codon of each CDS (Kumar et al., 2015). Such a widespread presence of large numbers of uORFs has serious implications for translation. The scanning model of translation initiation in eukaryotes proposes that the ribosome recognizes the 5′ cap and moves along with the transcript until the start codon of the CDS is reached (Kozak, 1978; Aitken and Lorsch, 2012; Hinnebusch et al., 2016). Scanning ribosomes will surely encounter numerous uORFs in the transcripts of *P. falciparum* before reaching downstream ORFs (dORFs), which are the protein-coding sequences (CDS) that encode the multitude of proteins required for the parasite’s complex life cycle.

As is the case with other eukaryotes (Child et al., 1999; Hood et al., 2009; von Arnim et al., 2014; Johnstone et al., 2016; McGee and Ingolia, 2016), uORFs in *P. falciparum* can...
downregulate the expression of the dORF (Amulic et al., 2009; Kumar et al., 2015). This post-transcriptional gene regulation (PTGR) mediated by uORFs takes place by engaging or stalling the ribosome at the uORF, subsequently decreasing the probability of the ribosome to initiate translation of the dORF (Morris and Geballe, 2000). It is remarkable that despite the frequent occurrence of uORFs in the majority of transcripts of *P. falciparum*, ribosomes are still able to translate the coding sequences present downstream (Lasonder et al., 2002). One mechanism through which the ribosome can reach the start codon of the dORF is leaky scanning where the ribosome skips the start codon of the uORF, continues scanning and reaches the AUG of the dORF to start translation (Kozak, 1984; Liu et al., 1984). Another mechanism employed by eukaryotic ribosomes to reach the dORF is reinitiation, where the ribosome translates the uORF and then reinitiates translation at the downstream AUG (Hughes et al., 1984; Kozak, 1984; Hinnebusch et al., 2016; Johnstone et al., 2016). The mechanism employed by the ribosome to reach the dORF is dictated by sequence features of the mRNA leader such as length of the uORF and the inter-cistronic length, which have been discussed in detail in this study.

Considering a large number of upstream AUGs (uAUGs) and uORFs present in *P. falciparum* mRNAs, it is of interest to understand how ribosomes are able to reach the dORF. In this report, we study the stages of the intra-erythrocytic developmental cycle (IDC) to investigate the effect of varying three features viz. uORF length, inter-cistronic length, and Kozak sequence on the translation of a dORF in *P. falciparum*. Codon usage is also expected to affect the ability of a uORF to influence translation at the dORF, however, as codon usage tables are available for *P. falciparum*, it is expected that the majority of transcripts contain multiple uORFs, yet the translation of the CDS still takes place. This work brings new insights into the mechanisms of cytoplasmic translation of mRNAs during the asexual life cycle of *P. falciparum*.

**Materials and methods**

**Mutational analysis of the Kozak sequence of the reporter gene**

Plasmid Pf86 (a kind gift from Kevin Milletello and Dyann Wirth, Harvard School of Public Health, Boston) contains a firefly luciferase reporter gene which is flanked by the 5′ leader sequence and 3′ untranslated region (UTR) of the gene for *P. falciparum* heat shock protein 86 (*Pfhs86*; Pf3D7_0708040) (Milletello et al., 2004). In this plasmid, 1837 base pairs of DNA (comprising the promoter region and a leader sequence of 686 base pairs from *Pfhs86*) are cloned upstream of the firefly luciferase coding region. To mutate the Kozak sequence of luciferase, the plasmid was treated with BstBI restriction enzyme (Thermo Fisher Scientific) whose recognition sites are present 16 nucleotides upstream and 165 nucleotides downstream of the start codon of a luciferase reporter gene. This fragment was replaced by a fragment which had a different Kozak sequence. DNA fragments with different Kozak sequences were generated by site-directed mutagenesis (SDM) of plasmid Pf86 using a set of forward primers to introduce desired Kozak sequences and a common reverse primer complementary to the ~165 nucleotide region. The primers used for this cloning are included in Additional File 1. The polymerase chain reaction (PCR) products were then digested with BstBI and ligated to BstBI-digested plasmid Pf86. The resulting colonies were screened for the presence of an insert with the desired mutation by restriction digestion as well as colony PCR. Clones were confirmed by sequencing.

**Generating clones for recombinant firefly luciferase, expressed in bacteria**

Wild type firefly luciferase had ‘G’ at its +4 position. To generate mutants of the gene with ‘A’/’T’/’C’ at +4 positions, SDM was performed with PCR using a forward primer having the desired mutation and a reverse primer complementary to the end of the CDS. Sequences of the primers are included in Additional File 1. This PCR fragment and the vector pET43a were separately digested with Xhol and Ndel restriction enzymes (Thermo Fisher Scientific). The insert fragment was ligated in the digested vector and transformed into *Escherichia coli* DH5α. Colonies were screened by digestion and clones were confirmed by sequencing. The resulting plasmid had a luciferase gene with mutations at the +4 position driven by the lac promoter and a C-terminal 6-Histidine tag for purification.

**Induction and purification of variant forms of firefly luciferase**

pET43a plasmids containing the firefly luciferase coding sequences with ‘A’, ‘T’, ‘G’ and ‘C’ at +4 positions were transformed into *E. coli* BL21 competent cells. To obtain variants of luciferase protein, the secondary culture of each variant was induced with 0.25 mM isopropyl β-D-1-thioglactopyranoside (IPTG). This was followed by 16 hours of incubation at 18 °C under shaking conditions until the OD600 reached 0.6. The harvested cells were re-suspended in 15 mL of ice-cold binding buffer (50 mM NaH2PO4, 300 mM NaCl and 10 mM imidazole). The cells were lysed by sonication with the cell lysis buffer and the luciferase proteins were purified with a protocol of Ni-NTA Superflow Cartridge (Qiagen) and the proteins were purified and analyzed on sodium dodecyl sulphate – polyacrylamide gel electrophoresis (SDS-PAGE) gels to check the purity of the enzymes.

**Luciferase assay of recombinant luciferase proteins**

Protein concentrations of purified luciferase enzyme variants (+4G, +4C, +4A and +4T) were quantified using the bicinchoninic acid (BCA) kit by following the manufacturer’s protocol (Sigma–Aldrich). An equal amount of each luciferase variant was diluted in Passive Lysis Buffer (Promega) and freeze-thawed for three cycles in liquid nitrogen to mimic the conditions of the luciferase assay done with parasites. The assay was performed using 90 μL of LAR reagent (Promega) and 10 μL of diluted proteins. Readings were captured for 30 s using a luminometer (Berthold Junior LB 9509).

**Removal of the native uORF from the Pf86 plasmid**

Plasmid Pf86 was used to test the effect of synthetic uORFs on the expression of the firefly luciferase reporter gene. The native uORF present 474 bases upstream in the 5′ leader was removed by

---

*euparasite*
mutating the start codon (ATG) to TTG by SDM to get Pf86*, a plasmid devoid of any uORF in the 5' leader. Primers used for this cloning are included in Additional File 1.

**Site directed mutagenesis to introduce uORFs**

SDM was used to introduce different uORFs in 5' leader sequence of Pfhsps86 cloned in plasmid PB86*. Specific regions in the 5' leader sequence were chosen for the introduction of the uORFs. This choice was dictated by the sequence of the 5' leader since the presence of AT-rich repeats resulted in technical problems with primer-binding during the procedure of SDM. Hence, the regions 191, 40 and 29 nucleotides upstream of the start codon of the luciferase reporter gene were chosen, as they had enough GC content to allow for annealing of the SDM primers. Non-overlapping forward and reverse primers containing the desired mutation were phosphorylated by polynucleotide kinase (New England Biolabs) as per the manufacturer’s protocol. PCR was carried out by Q5 HiFi DNA polymerase (New England Biolabs). Primers used for SDM are included in Additional File 1. Purified PCR products were treated with FastDigest DpnI restriction enzyme (Thermo Fisher Scientific) to eliminate the parental plasmid vector. Linear PCR products were circularized by ligation with T4 DNA ligase (Thermo Fisher Scientific). The final product was transformed into competent E. coli DH5α. The clones were confirmed to contain the desired mutation by sequencing.

**Increasing the length of uORF4 by introducing repeating units**

The sequence of uORF4 was designed in a way that it contained a recognition site for restriction enzyme AvrII (Thermo Fisher Scientific). The length of this uORF was thus increased by digesting the plasmid uORF4-191 and ligating it to annealed oligonucleotides of increasing length. In this process, the recognition site was lost after ligation. Screening was done by restriction digestion of the plasmid uORF4-191 and ligating it to annealed oligonucleotides in this study for more accuracy (Caro et al., 2014). A Python script was written to extract uORFs from the 5' leader sequences of 3137 transcripts that are expressed in asexual blood stages. The nucleotide at the +4 position of each uORF, the length of the uORF, and the inter-cistronic length from the CDS were calculated. The list of uORFs with these features is given in Additional File 2. Categorization of uORFs based on different features (nucleotide at +4 position, length and inter-cistronic length) was done. GO term enrichment analysis was done by using the GO enrichment tool of PlasmoDB (Release 41) (Aurrecoechea et al., 2009).

**Codon Adaptability Index (CAI) calculations**

The CAI of all the uORFs used in this study was calculated by using the algorithm described previously (Sharp and Li, 1987). The frequency of each codon in *P. falciparum* was obtained from previously published work (Nakamura et al., 2000).

**Results**

**The +4 position of the Kozak sequence plays a major role in translation initiation**

Previous work indicated that the +4 position of the Kozak sequence plays a critical role in translation initiation while the nucleotides preceding the start codon (−5 to −1) have no significant contribution towards the strength of the Kozak sequence in *P. falciparum* (Kumar et al., 2015). In this report, the effect of a ‘T’ at the +4 position could not be determined as changing the ‘G’ at the +4 position to ‘T’ in the luciferase expression vector (PB86, designated as control), altered the wild type codon ‘GAA’ to a stop codon. To address this lacuna, we decided to change the +5 position to ‘C’ in PB86 such that now a ‘T’ at the +4 position would generate a TCA codon. In order to facilitate a direct comparison to the data reported for ‘A’, ‘C’ and ‘G’ at the +4 position, the experimental approach taken in this section was identical to the one described in the study by Kumar et al. (2015).

The control plasmid PB86 was modified to generate seven constructs where the luciferase start codon was surrounded by Kozak sequences all of which had ‘T’ at the +4 position and different sequences from the −5 to −1 positions. These were transiently transfected into *P. falciparum* 3D7 parasites along with 50μg of plasmid containing Renilla luciferase as a control for transfection efficiency. The activity of the reporter gene in each of these mutants was measured and compared with that of wild type PB86. Attempts to quantify luciferase transcript levels were
unsuccessful due to the low transfection efficiency that has been reported for *P. falciparum* (Deitsch et al., 2001; Skinner-Adams et al., 2003; Hasenkamp et al., 2012; Rug and Maier, 2013). Therefore, the results shown in this section have been described as a combined effect of variation in transcription and translation of the luciferase reporter gene owing to the changes introduced in the Kozak sequence; hence, the word 'expression' is used.

The expression of constructs with variant Kozak sequence with 'T' at the +4 position was assayed, it was observed that despite differences in the −5 to −1 positions, the luciferase expression for each construct with 'T' at +4 position was ~60% of the control (Fig. 1A). Consistent with the earlier study, the strength of the different Kozak sequences having 'T' at the +4 position and varying nucleotides at the −5 to −1 positions, does not correlate with the frequency of each Kozak sequence in the genome of *P. falciparum* (Fig. 1A).

It is important to note that changing the +4 position changes the second amino acid of the reporter gene from wild type glutamate (GAA) to glutamine (CAA), lysine (AAA) or serine (TCA). The amino acid differences at the second position might have effects on luciferase reporter activity, as deleting the first seven amino acid residues from firefly luciferase resulted in a significant decrease in the enzyme activity of the protein (Wang et al., 2002).

If this was the case for the mutants reported here, the altered luciferase activities would be the outcome of altered luciferase enzyme, rather than alterations in the expression levels of the enzyme due to the Kozak sequence.

In order to eliminate this possibility, the four variants of the luciferase gene (corresponding to the luciferase proteins generated by making mutations at the +4 position) were cloned into a PET43a expression vector and expressed as recombinant proteins (Fig. 1B). The activities of these recombinant luciferase enzyme variants were compared to the wild type luciferase enzyme that has glutamate immediately following the start methionine. It was found that changing the second amino acid from glutamate (GAA) to glutamine (CAA), lysine (AAA) or serine (TCA) in the firefly luciferase enzyme did not change the luciferase activity significantly (Single-factor ANOVA test; \( P = 0.1026 \)) (Fig. 1C). Extrapolating from this data, luciferase enzyme activity in parasites transfected with constructs containing each of the four nucleotides at the +4 position of the Kozak sequence in Pf86, would not result from non-functional enzymes and instead, could be explained by changes in translation initiation driven by altered strengths of the *P. falciparum* Kozak sequence.

By incorporating the data shown here for 'T' at the +4 position with the data published by Kumar et al. (2015), the average firefly luminescence readings obtained for different Kozak sequences were grouped according to the nucleotide at the +4 position (Fig. 1D). Constructs that had Kozak sequences with 'G' at the +4 position showed the highest reporter gene activity (~95% as compared with the wild type variant), while constructs containing Kozak sequences with 'A' at the +4 position show the lowest activity (~8%) of the reporter gene. The constructs with 'T' or 'C' at the +4 positions showed intermediate reporter gene activities of ~64% and 30%, respectively.

These data lead us to assess the +4 positions of coding sequences (CDS) and uORFs in the *P. falciparum* 3D7 genome (PlasmoDB v24), and using bioinformatics, the sequences of all uORFs were extracted from 5' leader sequences of 3137 genes expressed in different stages of the IDC (Additional File 2). Data for the 5' leader lengths was exported from a previous study which predicted the length of leader sequences using RNA sequencing of ribosome footprints in RNAs isolated from asexual cultures (Caro et al., 2014).

Out of 3137 transcripts, only 67 contained no uORF. On the contrary, 3070 transcripts (i.e. greater than 97% of the transcripts analysed) contained at least one uORF, as reported earlier (Kumar et al., 2015). A total of 36769 uORFs were predicted from 3070 transcripts. To understand the distribution of uORFs across the genome, the number of uORFs was plotted against the number of CDS (Fig. 2A). A large number of transcripts have uORFs in the range of 6–10 with an average of 11 uORF per CDS. However, the transcript of a gene PF3D7_1107800 (AP2 domain transcription factor, putative) contains 123 uORFs, which is the highest number of uORFs reported for any gene in *P. falciparum* (Additional File 2). It was also observed that the number of uORFs is strongly correlated with the length of the leader sequence with a high R-square value of 0.8144 (Fig. 2B). These data suggest that uORFs are dispersed along the 5' leader lengths in *P. falciparum* with longer leader sequences (Watanabe et al., 2002; Caro et al., 2014) having a higher probability of containing uORFs.

Having extracted uORFs from the genome, analysis of the nucleotide at the +4 position of their Kozak sequences was performed.

The frequency of finding each of the four nucleotides at the +4 position of the Kozak sequences of annotated CDS and uORFs was computed (Fig. 3A and B). This genome-wide analysis revealed that CDS are most likely to have 'A' at the +4 position (47%) while 'G' follows next, being found in 30% of the CDS. The proportion of CDS with 'T' and 'C' at their +4 position is 16% and 7%, respectively. On the other hand, only 8% of uORFs had 'G' at the +4 position, and 9% of uORFs had 'C' at the +4 position. As expected, 'T' was seen in the +4 position of 48% of the uORFs and 'A' was found in 35%, consistent with the AT-bias of intergenic regions (>90% AT) being higher than that of the CDS (~60–70% AT).

Given that the 'G' at the +4 position resulted in a strong Kozak sequence, it is likely that 92% of uORFs have lower probabilities of engaging the ribosome. However, 8% of the total uORFs present in the *P. falciparum* genome have a 'G' at the +4 position, indicating that ~2900 uORFs engage the ribosome with high probability and have the potential to repress expression of the downstream CDS. These uORFs may result in lower translation efficiency of the doRFs with which they are associated or may play regulatory roles in translation. Alternatively, although reinitiation is not commonly observed in model eukaryotes, in *P. falciparum*, these data suggest that thousands of uORFs may have properties that allow the ribosome to reinitiate at the doRF.

Reinitiation is suggested by another observation from this analysis. Interestingly, 56% of the 36769 uORFs predicted in 3070 protein-coding transcripts have 'G' or 'T' at the +4 position (Fig. 3B), and we show that these types of features yield a stronger Kozak sequence as compared to the ones which have 'A' and 'C' at their +4 position (Fig. 1D). This is consistent with a previous study, which reports that approximately half of the total ribosome footprint coverage in 5' leader sequences of mRNAs overlaps with predicted uORFs (Caro et al., 2014).

In an attempt to find categories of genes with uORFs having different Kozak sequences, GO term analysis was performed with a P value cut off of 0.05 using the Gene Ontology Environment tool in PlasmoDB Release 41 (Aurrecoechea et al., 2009). GO terms were enriched for gene sets that are associated with uORFs having Kozak sequences with different nucleotides at their +4 position (Additional File 3). The terms that were common between all four sets had high ~ log10(P value) in the range of 15 and were associated with DNA replication, translation, response to stress, cellular transport, and localization (data not shown). This indicates that uORFs with different nucleotides at +4 position are distributed across different classes of genes.

However, classes of genes having GO terms that were different in each set were seen at lower values of ~ log10(P value). These classes of genes are shown in the radar chart according to their P value from Fisher's exact test (Fig. 3C). Genes involved in tightly
Fig. 1. The effect of changing the +4 position of the Kozak sequence on the expression of the reporter gene. (a) Firefly luciferase readings for constructs that contain the luciferase reporter gene with different Kozak sequences, as compared to the luciferase readings obtained from the control (wild type Pf86). The start codon of the reporter gene in different constructs is surrounded with the Kozak sequences having 'T' at +4 position as shown in the X-axis of the bar graph. The frequency of occurrence of the Kozak sequence in the genome of *P. falciparum* is plotted on the secondary Y-axis as points. The firefly luminescence units were normalized against those of Renilla luciferase for each construct. The firefly luciferase readings of each construct are given with respect to normalized firefly readings of the control (Pf86 with Kozak sequence CGGCCatgG). The readings and the standard deviations have been calculated from three replicates. The normalized firefly readings of Pf86 were in the range of 3000 to 6000 RLUs, measured in Berthold luminometer. (b) Recombinant luciferase enzyme variants with different amino acid at the second position were purified and electrophoresed on SDS-PAGE. (c) Activity of equal amount of recombinant luciferase variants was tested and Relative firefly Luminescence Units (RLUs) recorded from four replicates. Single-factor ANOVA test; *P* = 0.1026. (d) Table summarizing the average firefly luminescence units obtained from constructs with different nucleotides at +4 position in the Kozak sequence. The values for 'G', 'C', and 'A' at +4 position have been calculated by averaging the percentage firefly luminescence units obtained from constructs in each set as reported in Kumar et al. (2015).
regulated processes such as protein sumoylation and methylation of macromolecules including histones, DNA, tRNA and proteins, emerged in the set of genes associated with uORFs having ‘G’ at their +4 position. Apart from these broad categories, specific genes such as the translocon component PTEX88/150, involved in translocation of proteins to the host RBC surface, were also seen in this set. Interestingly, an analysis of all CDS in the genome, regardless of whether the gene is transcribed during the IDC, showed a large enrichment of genes involved in host–parasite interactions having ‘G’ at the +4 position of their Kozak sequences. These genes include the var, rifin and stevor genes that play roles in antigenic variation, suggesting that these gene families may be under translational control. Indeed, a member of the var family, the var2csa gene, also has a uORF with a Kozak sequence having ‘G’ at the +4 position and this uORF plays a role translational repression of the dORF (Kozak, 1987; Luukkonen et al., 1995; Child et al., 1999) and similar results have been reported for the var2csa gene in P. falciparum (Bancells and Deitsch, 2013). It has been proposed that the long inter-cistronic lengths allow the ribosomes to reacquire the ternary complex thereby increasing the probability of reinitiation.

Therefore, the effect of changing the inter-cistronic length on expression of the dORF was tested. Here, the 5′ leader sequence of the hsp86 gene cloned upstream of a luciferase reporter, was analyzed in the plasmid Pf86. The 5′ leader sequence has a native ORF (7 amino acids in length), 474 bases upstream of the luciferase start codon. This uORF was termed native-474 and its translatability score based on the Codon Adaptability Index (CAI) (Sharp and Li, 1987) was computed to be 0.759. This uORF has a "T" at the +4 position of the Kozak sequence, a nucleotide that should give an intermediate Kozak strength (~65% that of ‘G’ at the +4 position).

The start codon of the native ORF was mutated in order to generate a construct which does not have any uAUG/uORF (termed Pf86*). In the subsequent experiments, Pf86* was taken as a control and activity of luciferase obtained by this construct was taken as 100%. Activity from all the subsequent constructs has been shown as a percentage of the control. The constructs were transiently transfectioned into P. falciparum 3D7 parasites along with plasmid containing Renilla luciferase as an internal control of transfection. Attempts to quantify luciferase transcript levels did not yield results because of the characteristically low transcription efficiency observed for P. falciparum (Deitsch et al., 2001; Skinner-Adams et al., 2003; Hasenkamp et al., 2012; Rug and Maier, 2013). Therefore, in the results shown in this section and the subsequent sections, the term ‘expression’ has been used to describe a combined effect of variation in transcription and translation of the reporter gene due to the changes introduced in the leader sequence.

When the expression of luciferase from Pf86* was compared to that from the plasmid containing native-474, it was seen that the native-474 construct gave ~40% of the luciferase activity of Pf86*. Therefore, mutating the start codon of the native ORF led to a 2.5-fold increase in the expression of the reporter gene (Fig. 4a). This result is consistent with a previous study which reports that even one uAUG/uORF can repress expression of dORF (Kumar et al., 2015). To test the effect of changing inter-cistronic distance, the native uORF was moved to a position corresponding to 29 bases upstream of the luciferase start codon (native-29). Decreasing the inter-cistronic length to 29 nucleotides gave ~20% luciferase expression (Fig. 4a) compared to Pf86*.

Another uORF, uORF1 (a synthetic uORF with a high CAI value of 1), was introduced at either 474 or 29 nucleotides upstream of the start codon of luciferase. For these constructs, uORF1-474 and uORF1-29 respectively, luciferase expression reduced to 55% and 12% of the Pf86* control (Fig. 4b). The +4 position for the Kozak sequence of uORF1 was ‘C’, a nucleotide that shows weaker ability to engage the ribosome (~30% that of ‘G’ at the +4 position). Both the uORFs tested, native uORF and uORF1, showed approximately equal levels of repression suggesting that repression depends on a combination of Kozak sequences and CAI scores.

**Decreasing the inter-cistronic length leads to increased repression of reporter gene expression**

The presence of 8% of uORFs having strong Kozak sequences with a ‘G’ at the +4 position suggested that in addition to leaky scanning, the translation machinery of P. falciparum might also employ reinitiation to reach the dORF. For reinitiation to occur, after translation termination at the uORF, a fraction of ribosomes would remain associated with the mRNA (Morris and Geballe, 2000). Successful reinitiation at the dORF would depend on the probability of these ribosomes regaining the ternary complex (eIF2, GTP, and initiator Met-tRNAiMet). In mammalian cells, longer inter-cistronic lengths between the uORF and the dORF lead to efficient reinitiation of the dORF (Kozak, 1987; Luukkonen et al., 1995; Child et al., 1999) and similar results have been obtained for the var2csa gene in P. falciparum (Bancells and Deitsch, 2013). It has been proposed that the long inter-cistronic lengths allow the ribosomes to reacquire the ternary complex thereby increasing the probability of reinitiation.
Finally, a short upstream ORF, uORF2 (a synthetic uORF with a high CAI value of 1), coding for two amino acids and having a 'T' at the +4 position of the Kozak sequence, was introduced in the 5' leader. Despite having a Kozak sequence of a strength corresponding to 64% of that of the strongest Kozak tested, the introduction of uORF2 191 nucleotides upstream of the start codon (uORF2-191) did not reduce luciferase expression. The combination of a strong Kozak sequence with a short coding sequence (uORF2-191) did not reduce luciferase expression. The combination of uORF2 191 nucleotides upstream of the start codon (uORF2-191) did not reduce luciferase expression. The combination of uORF2 191 nucleotides upstream of the start codon (uORF2-191) did not reduce luciferase expression. From our data, it can be concluded that introducing a uORF close to the start codon of dORF leads to a significant decrease in the expression of the dORF. In all the cases, the maximal reduction in reporter activity was seen when the uORF was introduced within 50 nucleotides upstream of the start codon of luciferase, irrespective of differences in CAI, Kozak sequence and length of the different uORFs. This is similar to mammalian cells, where a decrease in the inter-cistronic length leads to inefficient translation of the dORF (Kozak, 1987; Luukkonen et al., 1995). Importantly, as inter-cistronic length has no effect on leaky scanning, the data are suggestive of reinitiation being the mechanism for translation of dORFs in P. falciparum when uORFs can engage the scanning ribosomes.

Driven by the experimental results of inter-cistronic lengths, a bioinformatics analysis of the genome of P. falciparum was undertaken to extract the sequences of uORFs that are present at different inter-cistronic lengths from the annotated CDS. Since uORFs that are closer to the CDS are expected to contribute more towards repression, all uORFs nearest to the CDS were extracted and the nucleotides at the +4 position of their Kozak sequence plotted against -log10 of P value from Fisher's exact test (Additional File 3). Only the GO terms that are unique for each nucleotide set have been plotted in the radar chart.

From our data, it can be concluded that introducing a uORF close to the start codon of dORF leads to a significant decrease in the expression of the dORF. In all the cases, the maximal reduction in reporter activity was seen when the uORF was introduced within 50 nucleotides upstream of the start codon of luciferase, irrespective of differences in CAI, Kozak sequence and length of the different uORFs. This is similar to mammalian cells, where a decrease in the inter-cistronic length leads to inefficient translation of the dORF (Kozak, 1987; Luukkonen et al., 1995). Importantly, as inter-cistronic length has no effect on leaky scanning, the data are suggestive of reinitiation being the mechanism for translation of dORFs in P. falciparum when uORFs can engage the scanning ribosomes.

Driven by the experimental results of inter-cistronic lengths, a bioinformatics analysis of the genome of P. falciparum was undertaken to extract the sequences of uORFs that are present at different inter-cistronic lengths from the annotated CDS. Since uORFs that are closer to the CDS are expected to contribute more towards repression, all uORFs nearest to the CDS were extracted and the nucleotides at the +4 position of their Kozak sequence plotted against -log10 of P value from Fisher's exact test (Additional File 3). Only the GO terms that are unique for each nucleotide set have been plotted in the radar chart.

From our data, it can be concluded that introducing a uORF close to the start codon of dORF leads to a significant decrease in the expression of the dORF. In all the cases, the maximal reduction in reporter activity was seen when the uORF was introduced within 50 nucleotides upstream of the start codon of luciferase, irrespective of differences in CAI, Kozak sequence and length of the different uORFs. This is similar to mammalian cells, where a decrease in the inter-cistronic length leads to inefficient translation of the dORF (Kozak, 1987; Luukkonen et al., 1995). Importantly, as inter-cistronic length has no effect on leaky scanning, the data are suggestive of reinitiation being the mechanism for translation of dORFs in P. falciparum when uORFs can engage the scanning ribosomes.

Driven by the experimental results of inter-cistronic lengths, a bioinformatics analysis of the genome of P. falciparum was undertaken to extract the sequences of uORFs that are present at different inter-cistronic lengths from the annotated CDS. Since uORFs that are closer to the CDS are expected to contribute more towards repression, all uORFs nearest to the CDS were extracted and the nucleotides at the +4 position of their Kozak sequence plotted against -log10 of P value from Fisher's exact test (Additional File 3). Only the GO terms that are unique for each nucleotide set have been plotted in the radar chart.

From our data, it can be concluded that introducing a uORF close to the start codon of dORF leads to a significant decrease in the expression of the dORF. In all the cases, the maximal reduction in reporter activity was seen when the uORF was introduced within 50 nucleotides upstream of the start codon of luciferase, irrespective of differences in CAI, Kozak sequence and length of the different uORFs. This is similar to mammalian cells, where a decrease in the inter-cistronic length leads to inefficient translation of the dORF (Kozak, 1987; Luukkonen et al., 1995). Importantly, as inter-cistronic length has no effect on leaky scanning, the data are suggestive of reinitiation being the mechanism for translation of dORFs in P. falciparum when uORFs can engage the scanning ribosomes.

Driven by the experimental results of inter-cistronic lengths, a bioinformatics analysis of the genome of P. falciparum was undertaken to extract the sequences of uORFs that are present at different inter-cistronic lengths from the annotated CDS. Since uORFs that are closer to the CDS are expected to contribute more towards repression, all uORFs nearest to the CDS were extracted and the nucleotides at the +4 position of their Kozak sequence plotted against -log10 of P value from Fisher's exact test (Additional File 3). Only the GO terms that are unique for each nucleotide set have been plotted in the radar chart.

From our data, it can be concluded that introducing a uORF close to the start codon of dORF leads to a significant decrease in the expression of the dORF. In all the cases, the maximal reduction in reporter activity was seen when the uORF was introduced within 50 nucleotides upstream of the start codon of luciferase, irrespective of differences in CAI, Kozak sequence and length of the different uORFs. This is similar to mammalian cells, where a decrease in the inter-cistronic length leads to inefficient translation of the dORF (Kozak, 1987; Luukkonen et al., 1995). Importantly, as inter-cistronic length has no effect on leaky scanning, the data are suggestive of reinitiation being the mechanism for translation of dORFs in P. falciparum when uORFs can engage the scanning ribosomes.

Driven by the experimental results of inter-cistronic lengths, a bioinformatics analysis of the genome of P. falciparum was undertaken to extract the sequences of uORFs that are present at different inter-cistronic lengths from the annotated CDS. Since uORFs that are closer to the CDS are expected to contribute more towards repression, all uORFs nearest to the CDS were extracted and the nucleotides at the +4 position of their Kozak sequence plotted against -log10 of P value from Fisher's exact test (Additional File 3). Only the GO terms that are unique for each nucleotide set have been plotted in the radar chart.

From our data, it can be concluded that introducing a uORF close to the start codon of dORF leads to a significant decrease in the expression of the dORF. In all the cases, the maximal reduction in reporter activity was seen when the uORF was introduced within 50 nucleotides upstream of the start codon of luciferase, irrespective of differences in CAI, Kozak sequence and length of the different uORFs. This is similar to mammalian cells, where a decrease in the inter-cistronic length leads to inefficient translation of the dORF (Kozak, 1987; Luukkonen et al., 1995). Importantly, as inter-cistronic length has no effect on leaky scanning, the data are suggestive of reinitiation being the mechanism for translation of dORFs in P. falciparum when uORFs can engage the scanning ribosomes.

Driven by the experimental results of inter-cistronic lengths, a bioinformatics analysis of the genome of P. falciparum was undertaken to extract the sequences of uORFs that are present at different inter-cistronic lengths from the annotated CDS. Since uORFs that are closer to the CDS are expected to contribute more towards repression, all uORFs nearest to the CDS were extracted and the nucleotides at the +4 position of their Kozak sequence plotted against -log10 of P value from Fisher's exact test (Additional File 3). Only the GO terms that are unique for each nucleotide set have been plotted in the radar chart.
Expression of the reporter gene depends on the length of uORF

So far, results shown in this report have indicated that in asexual stages of *P. falciparum*, reinitiation takes place to allow the ribosome to handle the large number of uORFs found in mRNAs. Reinitiation efficiency depends on the length of the uORF, with longer uORFs presumably resulting in the loss of initiation factors from the ribosome, in turn, decreasing initiation efficiency in mammalian cells (Kozak, 1987, 2001; Luukkonen et al., 1995; Hood et al., 2009). It is proposed that initiation factors involved in scanning, remain bound to the 40S subunit of the ribosome briefly after translation has been initiated. As the ribosome elongates a uORF with a longer length, these factors dissociate from the ribosome, thereby decreasing the probability of reinitiation at the dORF.

Possible effects of the length of the uORF were investigated in a bi-cistronic transcript by introducing ORFs of varying lengths upstream of the luciferase reporter gene. A 12 nucleotide long upstream ORF, uORF4 (CAI: 0.484) was introduced 191 nucleotides upstream of the start codon (uORF1-474 and uORF1-29, respectively). The firefly luciferase units obtained when a uORF present natively (Native-474) in the Pf86 plasmid 474 nucleotides upstream of the start codon was moved to 29 nucleotides upstream of the luciferase reporter gene (Native-29). The firefly luciferase units obtained when uORF1 was present at 474 and 29 nucleotides upstream of the start codon (uORF1-474 and uORF1-29, respectively). (C) The firefly luminescence units obtained when uORF2 was present at 191 and 40 nucleotides upstream of the start codon (uORF2-191 and uORF2-40 respectively). (D) The GO terms enriched in the genes associated with uORFs with the inter-cistronic length less than or equal to 50 nucleotides. The enriched GO terms are plotted against their $-\log_{10}$ of P-value from Fisher’s exact test (Additional File 3).

**Fig. 4.** The effect of inter-cistronic distance on the repression of reporter gene expression. Pf86*, a construct which does not harbour any uORF has been used as positive control while Pf86end, a construct in which the second codon of luciferase reporter gene is mutated to stop codon has been used as a negative control.
Increasing the length of upstream ORF, while keeping the Kozak sequence and inter-cistronic length constant, led to a gradual decrease in reporter gene expression. Although the CAI value of uORF4 was approximately half as compared to the other uORFs, the CAI values of uORF4-1X through uORF4-5X increased incrementally from 0.837 to 0.929. It is expected that increased CAI values would reflect uORFs that are easier to translate, and should result in a higher probability of the ribosome reaching the dORF and therefore, an increase in luciferase activity. Instead, luciferase activity showed a decrease in the presence of longer uORFs.

This suggests that as in other eukaryotes, in *P. falciparum*, apart from the Kozak sequence and inter-cistronic length, the length of the uORF also has a role to play in the expression of the dORF. These findings could be due to multiple reasons. For example, the addition of repeat elements to increase the length of the uORF could change the secondary structure of the mRNA which might affect the translation of the dORF. Nevertheless, given similar results in other eukaryotes (Kozak, 1987, 2001; Luukkonen et al., 1995; Hood et al., 2009), one of the mechanisms by which uORF length affects the expression of the dORF in *P. falciparum* could also be at the level of translation. In the presence of longer uORFs, the ribosomes have a lower probability of reaching the reporter gene and hence, the expression is repressed. In accordance with the data presented in the previous section, these results also suggest that reinitiation occurs in *P. falciparum*.

Here we show that uORFs coding for peptides longer than 19 amino acids (60 nucleotides) result in repression greater than 50% as compared to the control plasmid having no uORF (Pf86*). As the *P. falciparum* genome has a large number of uORFs, for the dORF to be translated, one would expect that the majority of uORFs should be less than 60 nucleotides long. As expected, the average length of predicted uORFs in the transcripts is 45 nucleotides (14 amino acids). The frequency distribution of the lengths of uORFs shows that 67% of ORFs are encoded by sequences of less than 45 nucleotides in length (Fig. 5b). The majority of these uORFs would allow the ribosome to reinitiate successfully at the start codon of the CDS. However, longer uORFs (length >45 nucleotides) may engage the ribosomes for a longer time and hence, reduce efficient reinitiation at the start codon of the dORF. A list of these uORFs and their downstream CDS has been given in the Additional File 2.
Differential GO term analysis between the sets of genes associated with uORFs whose length is less than and those greater than 45 nucleotides does not reveal enrichment of any category (data not shown). However, a set of genes that are associated with uORFs with length greater than 200 nucleotides was enriched in GO categories involved in transcription regulation, housekeeping processes (lipid catabolism and pH regulation) and protein ubiquitination and phosphorylation (Fig. 5c) (Additional File 3). Interestingly, the AP2 domain transcription factor family which is involved in transcriptional regulation during the development of P. falciparum (Painter et al., 2011) has numerous uORFs that are longer than 200 nucleotides. The uORFs present in the 5’ leaders of these genes might be involved in regulating the expression of this transcription factor and hence, controlling development in the parasite.

**Assessing the contribution of re-initiation in the expression of a dORF**

Results shown so far have indicated that P. falciparum parasites, despite having multiple uAUGs/uORFs in the 5’ leaders of their mRNAs, are able to translate their CDS by leaky scanning to bypass weak Kozak sequences and by re-initiation at the dORF. Reinitiation has been described previously in P. falciparum for translation of the var2csa gene in the presence of a 360 nucleotide long uORF (Bancells and Deitsch, 2013), with a parasite translation factor, PTEF (Chan et al., 2017), being induced in parasites that are found in the placenta of pregnant women suffering from malaria. Our data show that reinitiation may occur frequently in asexual parasites that are not sequestered to chondroitin sulphate in the placenta. Next, we analysed whether leaky scanning and reinitiation occur simultaneously in the asexual stages, in the presence of a single repressive uORF.

The strategy involved cloning a single uORF, out of frame with the dORF (luciferase reporter gene). The next step was to eliminate all the in-frame stop codons downstream of the uORF by mutation. To test the extent of reinitiation, the stop codon of the uORF is mutated. In case ribosomes rely on reinitiation alone as the mode of translation of the dORF, the expression of the dORF should be eliminated after mutating the stop codon. This is because the uORF is out of frame with the dORF and if initiation takes place, with no in-frame stop codons, the resulting protein will not encode luciferase. In case the ribosomes do not rely on reinitiation at all and instead use leaky scanning and other mechanisms, the expression of the dORF should remain the same as it was before mutating the stop codon. However, if a mix of reinitiation and other mechanisms, including leaky scanning occurs, the expression should be less than the expression seen before mutation of the stop codon of the uORF (Fig. 6a).

To assess the contribution of reinitiation in the translation of the dORF, the construct containing uORF2-191 (1T at the +4 position; 9 nucleotides in length), which does not repress the dORF (Fig. 4c) was selected. The presence of this uORF results in a high probability of the ribosome reaching the dORF either through reinitiation, leaky scanning or other mechanisms, thus providing an excellent scenario to test the contribution of reinitiation in the translation of the dORF.

The strategy involved cloning a single uORF, out of frame with the dORF (luciferase reporter gene). The next step was to eliminate all the in-frame stop codons downstream of the uORF by mutation. To test the extent of reinitiation, the stop codon of the uORF is mutated. In case ribosomes rely on reinitiation alone as the mode of translation of the dORF, the expression of the dORF should be eliminated after mutating the stop codon. This is because the uORF is out of frame with the dORF and if initiation takes place, with no in-frame stop codons, the resulting protein will not encode luciferase. In case the ribosomes do not rely on reinitiation at all and instead use leaky scanning and other mechanisms, the expression of the dORF should remain the same as it was before mutating the stop codon. However, if a mix of reinitiation and other mechanisms, including leaky scanning occurs, the expression should be less than the expression seen before mutation of the stop codon of the uORF (Fig. 6a).
would be due to a dependence on reinitiation. In the case of uORF2-191, the chances of reinitiating the dORF could be high due to 191 nucleotides in the inter-cistronic length and a small uORF length, paving the way for the ribosome to reacquire initiation factors.

Another uORF, uORF3-65 which is 135 nucleotides in length was introduced 65 nucleotides upstream of the luciferase start codon. The Kozak sequence of uORF2-191 and uORF3-65 are the same due to which the extent of leaky scanning in both the cases would be expected to be the same, however, the extent of reinitiation would differ for the two uORFs. As discussed in the previous sections, the length of this uORF and the short inter-cistronic distance would predict that the ribosomes have a lower probability of reacquiring initiation factors. Therefore, the contribution of reinitiation to the translation of the dORF should be less than that seen for uORF2-191. Consistent with these predictions, the luciferase expression obtained in presence of this uORF was 39% of that of Pf86* and the mutation of the stop codon of uORF3-65 led to a further, yet a small decrease in expression (25% of Pf86*) as compared to the case when the stop codon was present (Fig. 6b). The probability of ribosomes reaching the reporter gene via reinitiation was calculated using the same strategy as used for uORF2-191 and seen to be 0.36 (Fig. 6c). This result is consistent with the expectation that the extent of reinitiation from uORF3-65 should be less than that of uORF2-191 (probability of 0.36 compared to 0.76).

For these two uORFs, a mix of leaky scanning and reinitiation is seen for expression of the dORF. These results reinforce the observations that reinitiation may be widespread in asexual stages of *P. falciparum*, and not restricted to parasites isolated from pregnancy-associated malaria (PAM) samples.

**Discussion**

In this report, a systematic study of the features that contribute to repression by uORFs resulted in the observation that *P. falciparum* asexual stage parasites employ widespread leaky scanning and reinitiation to allow the scanning ribosome to reach the dORF and express a multitude of proteins. Some classes of proteins may undergo translational regulation as uORFs associated with their CDS have features associated with repression: strong Kozak sequences, inter-cistronic lengths less than 50 nucleotides and/or lengths greater than 200 nucleotides. Specifically, these classes of proteins include antigenic variation gene families, including var, rifin and stever, proteins involved in ubiquitination and members of the AP2 transcription factor family. Potential mechanisms by which translational regulation of these classes of proteins by uORFs might occur are discussed.

**Reinitiation of the downstream ORF is a strategy used by *P. falciparum* to handle large numbers of repressive uORFs**

We show that uORF length and inter-cistronic distance affect the expression of the downstream luciferase reporter gene, both suggesting that reinitiation takes place in *P. falciparum* in asexual stages, including the parasites that bind to the placenta as previously reported (Bancells and Deitsch, 2013). Data suggest that the ribosome is able to reinitiate translation of CDS when uORFs are less than 60 nucleotides long. This is comparable to other eukaryotes where uORFs that allow reinitiation are usually less than 30, 48 and 90 nucleotides in *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and mammals, respectively (Kozak, 2001; Hinnebusch, 2005; Calvo et al., 2009; von Arnim et al., 2014). In case of inter-cistronic length, we show that uORFs within 50 nucleotides of the start codon repress the expression of dORF drastically in *P. falciparum*. A similar observation has been made for viruses, where expression of the dORF decreases when the inter-cistronic length is reduced from 64 to 16 nucleotides (Luukkonen et al., 1995). Our analysis shows that 40% transcripts in the IDC stages of *P. falciparum* have at least one uORF within 50 bases upstream of start codon of CDS. We propose that to translate the CDS, *P. falciparum* would utilize molecular factors to enhance the probability of reinitiation.

One such player is PTEF (Plasmodium translation enhancing factor) which is required for reinitiation of translation of the *var2 CSA* CDS in the presence of a uORF (Chan et al., 2017). This 360-nucleotide long uORF has a ‘G’ at the +4 position of the Kozak sequence; both these features have been reported to cause repression in the expression of dORF in the present study, presumably by allowing the scanning of ribosomes to initiate at the strong Kozak sequence, allowing them to translate the uORF and terminate before reaching the var2 CSA CDS. In order to relieve this repression, PTEF, a factor that promotes reinitiation is required. PTEF is expressed in intra-erythrocytic stages with expression levels increasing from ring to schizont stage (Pease et al., 2013). However, in the case of pregnancy-associated malaria, 7–15-fold higher expression of PTEF appears to be needed for translation of *var2 CSA* due to the presence of the 360-nucleotide long uORF which is significantly longer than the average length of predicted uORFs (Chan et al., 2017). One might speculate that in asexual stage parasites, low levels of PTEF are sufficient to enable reinitiation in the presence of numerous smaller uORFs which are abundant in transcripts. The presence of wide-spread reinitiation, possibly driven by PTEF, leads to the speculation that as this protein shows low sequence identity with human proteins, it might be a potential drug target for the blood stages of *P. falciparum*.

Another molecular player that is involved in reinitiation is the translation initiation factor, eIF2α. Evidence of reinitiation mediated by phosphorylation of eIF2α during nutritional stress responses is well documented for the *S. cerevisiae* GCN4 transcript (Hinnebusch, 1993, 2005) and for the integrated stress response in mammalian cells (Young et al., 2015). In the case of *P. falciparum*, one would expect that if there is a significant amount of reinitiation, PteIF2a should be phosphorylated. Interestingly, phosphorylation of PteIF2α is not seen in ring and trophozoite stages, however, a sudden increase is observed in schizonts (Zhang et al., 2017). Although suggestive of reinitiation occurring predominantly in schizonts, due to a lag observed between the peaks of mRNA synthesis and corresponding protein abundance in *P. falciparum* (Le Roch et al., 2004; Foth et al., 2011), it is also possible that the phosphorylated PteIF2α might promote reinitiation from mRNAs expressed at earlier IDC stages. Another molecular factor that plays a role in reinitiation is eIF3, warranting further work on this protein in *P. falciparum*. Eukaryotic IF3 (eIF3) remains transiently attached to the elongating ribosome and stabilizes the post-termination complex to stimulate reinitiation of the dORF in *S. cerevisiae* (Cuchalová et al., 2010; Hronová et al., 2017; Mohammad et al., 2017). Similarly, the H subunit of eIF3 helps in an efficient resumption of scanning after the translation of the uORF in *Arabidopsis* (Roy et al., 2010).

**The Kozak sequence determines the extent of leaky scanning**

Alleviation of translational repression due to numerous uORFs can also be achieved through leaky scanning, possibly the most metabolically efficient way to handle multiple uORFs. A known contributor to leaky scanning is the strength of the Kozak sequence (Kozak, 1999; Ferreira et al., 2013, 2014). This report confirms published observations that the nucleotide following the start codon (+4 position) plays a significant role in determining the strength of the Kozak sequence in *P. falciparum* (Kumar
et al., 2015), unlike other eukaryotes where nucleotides at −3 and +4 positions are both important (Pisarev et al., 2006).

In addition to Kozak sequences, eIF1, a translation initiation factor in the preinitiation complex can facilitate selection of the start codon (Cheung et al., 2007) by the release of eIF1 from the preinitiation complex leading to translation initiation (Pestova et al., 1998; Passmore et al., 2007; Nanda et al., 2009). High concentrations of eIF1 lead to the stringent selection of AUGs having strong Kozak sequences (Loughran et al., 2012; Andreev et al., 2015; Fijalkowska et al., 2017) and the phosphorylation state of eIF1 under stress conditions also determines the selection of the start codon, helping to bypass start codons with weak Kozak sequences (Zach et al., 2014). Another factor, the m7G-cap-binding factor eIF4G1, also enhances leaky scanning of uORFs near the cap when bound to eIF1 in mammalian cells (Haimov et al., 2018). Yet another factor that affects leaky scanning in eukaryotes is eIF2α. When eIF2α is phosphorylated under stress conditions, uORFs with weak Kozak sequences are more likely to be bypassed by leaky scanning (Palam et al., 2011). Interestingly, eIF2α phosphorylation also facilitates reinitiation (Hinnebusch, 2005), suggesting that this protein could be a key player in handling repression by uORFs.

**Implications of uORFs in gene regulation**

Like other eukaryotes, *P. falciparum* utilizes a spectrum of mechanisms of gene regulation ranging from epigenetic (Caro et al., 2014; Saraf et al., 2016) to post-transcriptional gene regulation (Bunnik et al., 2013). However, unlike other eukaryotes, *P. falciparum* seems to lack the variety of canonical eukaryotic transcription factors or stage-specific transcription factors (Gardner et al., 2002; Coulson et al., 2004). Stage-specific regulation at the transcriptional level is achieved via the AP2 family of transcription factors (Balaji et al., 2005). This family of transcription factors is known to control the development of the parasite through various stages (Painter et al., 2011). Interestingly, transcript and protein levels during each stage of the parasite’s life-cycle are poorly correlated which points towards translationally controlled expression (Le Roch et al., 2004) and a translational regulator, PfALBA1, represses translation of transcripts involved in invasion until the parasite is mature enough to invade (Vembar et al., 2015, 2016).

Apart from PfALBA1, cis-regulatory factors such as uORFs downregulate the expression of dORFs. The frequent occurrence of uORFs in the transcripts of *P. falciparum* poses an additional layer of post-transcriptional gene regulation (Kumar et al., 2015). One way to regulate the number of uORFs is by changing in the leader length of the transcripts which has been observed during different stages (Caro et al., 2014). Another regulatory mechanism to remove uORFs could involve alternative splicing of the leader sequence. Evidence for this possibility lies in observations that 431 splice junctions fall in intergenic regions of genes (Sorber et al., 2011). For the remaining uORFs that still pose a hindrance to the ribosome, we propose that a mix of leaky scanning and reinitiation could regulate the expression of the dORF, not only during the IDC but also during the numerous situations of stress that are faced by the parasite during multiplication and development in the human host.

Widespread use of reinitiation and leaky scanning has been observed during stress conditions in a multitude of eukaryotes (Hinnebusch, 1993, 2005; von Arnim et al., 2014; Zach et al., 2014; Andreev et al., 2015; Young et al., 2015; Hinnebusch et al., 2016; Fijalkowska et al., 2017). Seemingly, organisms facing stress undergo non-canonical translation to economize their energy and resource usage. Hence, investigating translation initiation factors during stress conditions in *P. falciparum* would provide insights into the role of multiple uORFs harboured by the parasite’s transcripts.

In addition to regulation of translation during stress, an assortment of non-canonical translation mechanisms including reinitiation, leaky scanning, internal ribosome entry and ribosome shunting has been observed in viruses (Ryabova et al., 2006; de Bryne and Ohlmann, 2018). Given their small genome size, viruses have adopted multiple strategies to transcribe and translate their protein repertoire. In the case of *P. falciparum*, the constraint is not genome size, but the high AT content of the genome. Owing to this, mRNAs have numerous uORFs and uAUGs which the parasite’s translation machinery appears to handle by widespread use of non-canonical translation mechanisms.

In conclusion, this is the first report that systematically delineates the features of uORFs that affect translation of the downstream gene in *P. falciparum*. Additionally, we show that a mix of reinitiation and leaky scanning mechanisms are employed in asexual stages of *P. falciparum* to translate the dORF in the presence of upstream ORFs. Therefore, initiation factors such as PTEF, PfEF1, PfEF2 and PfEF3 may be critically involved in translation and regulation of parasite proteins. This distinguishing feature of the *P. falciparum* cytoplasmic translation machinery has the potential to become a novel target for anti-malarial drugs.

**Supplementary material.** The supplementary material for this article can be found at [https://doi.org/10.1017/S0031182020000840].

**Acknowledgements.** We thank IIT Bombay Hospital, staff, and volunteers for assisting in blood collection for parasite culturing.

**Author contribution.** CK, MK and SP conceived the study and designed the experimental setup. CK and MK carried out the experiments. CK and SP wrote the manuscript. All authors read and approved the final manuscript.

**Financial support.** This work was partially funded by intramural funds from IIT Bombay. CK is supported by a PhD Teaching-Assistant Fellowship from IIT Bombay. MK was supported by a Research Fellowship from the University Grants Commission (UGC).

**Conflict of interest.** The authors declare that they have no competing interests.

**Ethical standards.** Fresh RBCs were collected from healthy volunteers after approval from the Institutional Ethics Committee of IIT Bombay.

**References**

Atiken CE and Lorsch JR (2012) A mechanistic overview of translation initiation in eukaryotes. Nature Structural & Molecular Biology 19, 568–576.

Amulic B, Salanti A, Lavatsen T, Nielsen MA and Deitsch KW (2009) An upstream open reading frame controls translation of var2csa, a gene implicated in placental malaria. PLoS Pathogens 5, e1000256.

Andreev DE, O’Connor PBF, Zhdanov AV, Dmitriev RI, Skatsny IN, Papkovsky DB and Baranov PV (2015) Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. Genome Biology 16, 90.

Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nakay V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Treatman C and Wang H (2009) Plasmodb: a functional genomic database for malaria parasites. Nucleic Acids Research 37, D339–D343.

Balaji S, Babu MM, Iyer LM and Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Research 33, 3994–4006.

Bancells C and Deitsch KW (2013) A molecular switch in the efficiency of translation reinitiation controls expression of var2csa, a gene implicated in pregnancy-associated malaria: var2csa is expressed by translation reinitiation. Molecular Microbiology 90, 472–488.
Chaubey S, Kumar A, Singh D and Habib S (2005) The apicoplast of... Malaria Journal 11, 210.

Hinnebusch AG (1993) Gene-specific translational control of the yeast GCN4 gene by phosphorylation of eukaryotic initiation factor 2. Molecular Microbiology 10, 215–223.

Hinnebusch AG (2005) Translational regulation of GCN4 and the general amino acid control of yeast. Annual Review of Microbiology 59, 407–450.

Hinnebusch AG, Ivanov IP and Sonenberg N (2016) Translational control by 5′-translated regions of eukaryotic mRNAs. Science (New York, N.Y.) 352, 1413–1416.

Hood HM, Neafsey DE, Galagan J and Sachs MS (2009) Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. Annual Review of Microbiology 63, 385–409.

Hronová V, Mohammad MP, Wagner S, Pánek J, Gunisová S, Zeman J, Poncová K and Valášek LS (2017) Does eIF3 promote reinitiation after translation of short upstream ORFs also in mammalian cells? RNA biology 14, 1660–1667.

Hughes S, Mellstrom K, Kosik E, Tamanoff F and Brugge J (1984) Mutation of a termination codon affects initiation. Molecular and Cellular Biology 4, 1738–1746.

Jackson KE, Habib S, Frugier M, Hoen R, Khan S, Pham JS, Ribas de Pouplana L, Royo M, Santos MAS, Sharma A and Ralph SA (2011) Protein translation in Plasmodium Parasites. Trends in Parasitology 27, 467–476.

Johnstone TG, Bazzini AA and Giraldes AJ (2016) Upstream OREs are prevalent translational repressors in vertebrates. The EMBO Journal 35, 706–723.

Kozak M (1978) How do eukaryotic ribosomes select initiation regions in messenger RNA? Cell 15, 1109–1123.

Kozak M (1984) Selection of initiation sites by eukaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. Nucleic Acids Research 12, 3873–3893.

Kozak M (1987) Effects of intercistronic length on the efficiency of reinitiation by eukaryotic ribosomes. Molecular and Cellular Biology 7, 3438–3445.

Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. Gene 234, 187–208.

Kozak M (2001) Constraints on reinitiation of translation in mammals. Nucleic Acids Research 29, 5226–5232.

Kumar M, Srinivas V and Patankar S (2015) Upstream AUGs and upstream ORFs can regulate the downstream ORF in Plasmodium falciparum. Malaria Journal 14, 512.

Kyes SA, Christodoulou Z, Raza A, Horrocks P, Pinches R, Rowe JA and Newbold CI (2003) A well-conserved Plasmodium falciparum var gene shows an unusual stage-specific transcript pattern. Molecular and Cellular Biology 13, 1339–1348.

Lasonder E, Ishihama Y, Andersen JS, Vermunt AMW, Pain A, Sauerwein RW, Eling WMC, Hall N, Waters AP, Stunnenberg HG and Mann M (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 419, 537–542.

Lavstsen T, Salanti A, Jensen AT, Arnot DE and Theander TG (2003) Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malaria Journal 2, 27.

Le Roch KG, Johnson JR, Flores L, Zhou Y, Santosyona A, Grainger M, Yan SF, Williamson KG, Holder AA, Carucci DJ, Yates JR and Winzeler EA (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Research 14, 2308–2318.

Liu CC, Simonsen CC and Levinson AD (1984) Initiation of translation at internal AUG codons in mammalian cells. Nature 309, 82–85.

Loughran G, Sachs MS, Atkins JF and Ivanov IP (2012) Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. Nucleic Acids Research 40, 2898–2906.

Luukkonen BG, Tan W and Schwartz S (1995) Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance. Journal of Virology 69, 4086–4094.

McGeachy AM and Ingolia NT (2016) Starting too soon: upstream reading frames repress downstream translation. The EMBO journal 35, 699–700.
Militello KT, Dodge M, Bethke L and Wirth DF (2004) Identification of regulatory elements in the Plasmodium Falciparum genome. Molecular and Biochemical Parasitology 134, 75–88.

Mohammad MP, Munzarová Pondelícková V, Zeman J, Gunisová S and Valášek LS (2017) In vivo evidence that eIF3 stays bound to ribosomes elongating and terminating on short upstream ORFs to promote reinitiation. Nucleic Acids Research 45, 2658–2674.

Morris DR and Geballe AP (2000) Upstream open reading frames as regulators of mRNA translation. Molecular and Cellular Biology 20, 8635–8642.

Nakamura Y, Gojobori T and Ikemura T (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Research 28, 292.

Nanda JS, Cheung Y-N, Takacs JE, Martin-Marcos P, Saini AK, Hinnebusch AG and Lorsch JR (2009) eIF1 controls multiple steps in start codon recognition during eukaryotic translation initiation. Journal of Molecular Biology 394, 268–285.

Painter HJ, Campbell TL and Llinás M (2011) The Apicomplexan AP2 family: integral factors regulating Plasmodium development. Molecular and Biochemical Parasitology 176, 1–7.

Palam LR, Baird TD and Wek RC (2011) Phosphorylation of eIF2 facilitates ribosomal bypass of an inhibitory upstream ORF to enhance CHO PROC translation. The Journal of Biological Chemistry 286, 10939–10949.

Pasaje CFA, Cheung V, Kennedy K, Lim EE, Baell JB, Griffin MDW and Griffin MDW (2016) ORFpred: a machine learning program to identify translatable small open reading frames in intergenic regions of the Plasmodium Falciparum genome. Current Bioinformatics 11, 000–000.

Vembark MS, Macpherson CR, Simeiro O, Coppée J-Y and Scherf A (2015) The PHAβ1 RNA-binding protein is an important regulator of translational timing in Plasmodium falciparum blood stages. Genome Biology 16, 212.

Vembark SS, Droll D and Scherf A (2016) Translational regulation in blood stages of the malaria parasite Plasmodium spp.: systems-wide studies pave the way. Wiley Interdisciplinary Reviews. RNA 7, 772–792.

von Arnim AG, Jia Q and Vaughn JN (2014) Regulation of plant translation by upstream open reading frames. Plant Science: An International Journal of Experimental Plant Biology 214, 1–12.

Wang XC, Yang J, Huang W, He L, Yu JT, Lin QS, Li W and Zhou HM (2002) Effects of removal of the N-terminal amino acid residues on the activity and conformation of firefly luciferase. The International Journal of Biochemistry & Cell Biology 34, 983–991.

Watanabe J, Sasaki M, Suzuki Y and Sugano S (2002) Analysis of transcriptomes of human malaria parasite Plasmodium falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene 291, 105–113.

Wong W, Bai X, Brown A, Fernandez IS, Hanssen E, Condon M, Tan YH, Baum J and Scheres SHW (2014) Cryo-EM structure of the Plasmodium Falciparum 80S ribosome bound to the anti-protozoan drug emetine. eLife 3, e03080. doi: 10.7554/eLife.03080

Wong W, Bai X-C, Sleels BC, Triglia T, Brown A, Thompson JK, Jackson KE, Hanssen E, Marapana DS, Fernandez IS, Ralph SA, Cowman AF, Scheres SHW and Baum J (2017) Melchofire targets the Plasmodium Falciparum 80S ribosome to inhibit protein synthesis. Nature Microbiology 2, 17031.

World Malaria Report 2018 (2018) World Health Organization.

Ye Y, Liang Y, Yu Q, Hu L, Li H, Zhang Z and Xu X (2015) Analysis of human upstream open reading frames and impact on gene expression. Human Genetics 134, 605–612.

Young SK, Willy JA, Wu C, Sachs MS and Wek RC (2015) Ribosome reinitiation directs gene-specific translation and regulates the integrated stress response. The Journal of Biological Chemistry 290, 28257–28271.

Zach L, Braunstein I and Stanhill A (2014) Stress-induced start codon fidelity regulates arthritis-inducible regulatory particle-associated protein (AIRAP) translation. The Journal of Biological Chemistry 289, 20706–20716.

Zhang M, Gallego-Delgado J, Fernandez-Arias C, Waters NC, Rodriguez A, Tsuji M, Wek RC, Nussenzeig V and Sullivan WJ (2017) Inhibiting the Plasmodium eIF2α Kinase PK4 Prevents Artemisinin-Induced Latency. Cell Host & Microbe 22, 766–776.e4.