SCIENTIFIC REPORTS

OPEN

# Forces acting on codon bias in malaria parasites

I. Sinha[1,2] & C. J. Woodrow[1,2]

Received: 8 August 2017
Accepted: 16 October 2018
Published online: 29 October 2018

**Malaria parasite genomes have a range of codon biases, with *Plasmodium falciparum* one of the most AT-biased genomes known. We examined the make up of synonymous coding sites and stop codons in the core genomes of representative malaria parasites, showing first that local DNA context influences codon bias similarly across *P. falciparum*, *P. vivax* and *P. berghei*, with suppression of CpG dinucleotides and enhancement of CpC dinucleotides, both within and aross codons. Intense asexual phase gene expression in *P. falciparum* and *P. berghei* is associated with increased A3:G3 bias but reduced T3:C3 bias at 2-fold sites, consistent with adaptation of codons to tRNA pools and avoidance of wobble tRNA interactions that potentially slow down translation. In highly expressed genes, the A3:G3 ratio can exceed 30-fold while the T3:C3 ratio can be less than 1, according to the encoded amino acid and subsequent base. Lysine codons (AAA/G) show distinctive behaviour with substantially reduced A3:G3 bias in highly expressed genes, perhaps because of selection against frameshifting when the AAA codon is followed by another adenine. Intense expression is also associated with a strong bias towards TAA stop codons (found in 94% and 89% of highly expressed *P. falciparum* and *P. berghei* genes respectively) and a proportional rise in the TAAA stop 'tetranucleotide'. The presence of these expression-linked effects in the relatively AT-rich malaria parasite species adds weight to the suggestion that AT-richness in the *Plasmodium* genus might be a fitness adaptation. Potential explanations for the relative lack of codon bias in *P. vivax* include the distinct features of its lifecycle and its effective population size over evolutionary time.**

Codon bias, a generalised tendency to use codons non-randomly, occurs both between and within individual organisms[1]. From the time of its discovery, adaptive explanations have been proposed, with early studies on *E. coli* ribosomal protein genes leading to a hypothesis that bacterial codon usage was an adaptive genome strategy for optimal translational efficency and/or accuracy[2–4]. Studies within individual organisms indicate that codon bias can influence transcription[5], mRNA editing and stability[6], translation initiation and elongation[7,8], and post-translational modification[9]. Increased translational speed is often attributed to adaptation of codons to the tRNA pool[7,8,10] but codon adaptation may also reflect selection for avoidance of 'wobble' codon-tRNA interactions (which slow down translation[11]) and frameshift-prone sequences.

In contrast, explanations for underlying biases in base composition across entire genomes (AT- or GC-richness) have tended to revolve around 'neutral' mechanisms[12,13], particularly with smaller effective population sizes in which selection has a relatively smaller role. For example, although a number of functional explanations for the substantial differences in AT-richness among prokarya have been proposed, there is good evidence for a common mutational pressure towards low GC[14].

Studying codon bias in malaria parasites is of potential importance for several reasons. Malaria is the most important parasitic disease of humans, causing over 200 million cases per year with around half a million deaths[15]. A detailed understanding of malaria biology is considered vital for control and elimination of these parasites, and the complete genome sequences of the key malaria species *Plasmodium falciparum*[16] and *Plasmodium vivax*[17] have proved a foundation for a wide range of basic and applied studies relevant to development of new drugs and vaccines[18]. Furthermore, these sequences are now linked to an extensive range of genome-wide studies of transcription and translation using both microarray and next-generation sequencing approaches[19–24], providing powerful datasets for integrated analysis of the association between expression and codon bias both within and across these important human pathogens. Finally, malaria parasites have distinctive intrinsic codon usage properties. *P. falciparum* has one of the most AT-rich genomes known; over the entire genome A or T nucleotides constitute 80.6% of all bases[16], and 85% of synonymous positions[17] with non-synonymous positions also

[1]Mahidol-Oxford Tropical Medicine Research Unit (MORU), Mahidol University, Bangkok, Thailand. [2]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK. Correspondence and requests for materials should be addressed to C.J.W. (email: charlie.woodrow@ouh.nhs.uk)

| Organism | Type of data | Number of genes with data | Reference |
|---|---|---|---|
| *P. falciparum* | Microarray | 4555 | [19] |
| *P. falciparum* | RNA-seq | 4802 | [20] |
| *P. falciparum* | Ribosome profiling | 3605 | [24] |
| *P. vivax* | Microarray | 4400 | [21] |
| *P. vivax* | RNA-seq | 4981 (SMRU1) | [22] |
| *P. berghei* | RNA-seq | 4979 | [23] |

**Table 1.** Expression datasets used in the work.

affected[25]. The 'rodent' malarias (*P. berghei*, *chabaudi* and *yoelii*) are evolutionary distant, but nearly as AT-biased (%A + T = 75.6–77.4%) while the (A + T) genome composition of *P. vivax* is 62.4%[26].

Relatively few studies have examined how codon bias varies between genes (or different parts of genes) within individual malaria parasite species. Evidence of expression-associated codon bias has already been described for *P. falciparum* and attributed to translational selection favouring particular codon-tRNA interactions[27,28]. Nucleotide context also appears important, with suppression of CpG dinucleotides first reported 30 years ago in *P. falciparum*[29–31] where it is present on a genome-wide scale[32] and proposed to result from CpG methylation and subsequent deamination of methylcytosine to T, a common property of many eukaryotic organisms[33].

In order to explore how the various mutational and selective forces underlying codon bias operate within and between malaria parasites, we undertook a genome-wide analysis of synonymous sites and stop codons in three representative malaria parasites (*P. falciparum*, *P. vivax* and *P. berghei*), exploring the relationship between local DNA context and codon bias. Then we explored the influence of the level of gene expression in the asexual phase of the lifecycle. Given the greater expression-related effects observed for *P. falciparum* and *P. berghei*, we studied how these influences act in combination to influence codon bias in these species. Finally we consider the implications of the association between expression-related bias and AT-richness in the *Plasmodium* genus.

## Methods

### Genomic data.

Coding sequences were downloaded from Plasmodb with analysis undertaken in custom Python packages and in R. To aid comparison across species, *P. falciparum* 3D7 sequences were chosen that avoided multigene families, producing a set of 5055 coding genes corresponding to the 'core' genome, containing a total of approximately 2.6 million synonymous nucleotides i.e. just over 20% of total coding sequence[16]. Coding sequences orthologous to these genes (defined via PlasmodDB) were examined for *P. berghei* (4496 genes) and *P. vivax* (4712 genes).

### Asexual stage expression data.

Data for *P. falciparum*, *P. berghei* and *P. vivax* were obtained from published datasets (Table 1). RNA-Seq data[20,22,23] were used for the main analyses; microarray data[19,21] for *P. falciparum* and *P. vivax* were also checked for consistency. In addition the relationship between *P. falciparum* stop codon usage and RNA translation as assessed by ribosome profiling[24] was studied. For each dataset, genes were ranked by maximum expression level across the lifecycle and then stratified into three levels; I: <75th percentile, II: 75–97th percentile, III: >97th percentile.

For comparisons of proportions between categories the Chi-squared test was undertaken; p values are shown in Supplementary data.

The parameter for strength of selected codon usage bias, *S* (which is 'confounded' by effective population size) was calculated for the nine two-fold synonymous codons (e.g. TTT and TTC coding for phenylalanine) using the formula described by dos Reis and Wernisch[34,35]: $S = \ln(P_{hx}/(1\text{-}P_{hx})) - \ln(P_{ref}/(1\text{-}P_{ref}))$, where $P_{hx}$ is the observed frequency of the codon with relatively higher expression in top expression band III, and $P_{ref}$ is its observed frequency in expression band I (i.e. $P_{hx}$ is higher than $P_{ref}$).

## Results

### Context-dependent codon bias.

*Between codon effects.* The clearest view of the influence of neighbouring nucleotide context on codon bias can be obtained by examining how the third position in 2-fold or 4-fold synonymous codons varies according to the identity of the first nucleotide of the following codon (termed $N_1$)[8]. Because $N_1$ lies in the next codon, any patterns observed are unlikely to be due to translational selection (although other selective forces are possible).

*2-fold synonymous sites (9 amino acids).* According to the genetic code, the effect of neighbouring nucleotide context at 2-fold synonymous sites (third codon position) can be explored via the ratio of thymines:cytosines (T3:C3) (for the six pyrimidine-ending amino acid codon pairs) and adenines:guanines (A3:G3) (for the three purine-ending amino acid codon pairs) (Fig. 1A).

A clear association is present between $N_1$ context and T3:C3 ratio in all three studied species (Fig. 2, Supplementary Table 1). Consistent with previous evidence of lower than expected CpG sites[29,32], T3:C3 ratio is highest (and cytosines hence most rare) with $N_1$G. For example, in *P. falciparum*, the T3:C3 ratio is 8.6 with $N_1$G vs. 5.9 with $N_1$H (H = 'not G') while in *P. berghei* T3:C3 is 7.3 with $N_1$G vs. 4.4 with $N_1$H. Incidentally the relatively low T3:C3 for cysteine in *P. berghei* agrees with previous results[36].

It is also clear that T3:C3 is lowest (i.e. cytosines more common) when followed by another cytosine ($N_1$C). This effect is consistent across all amino acids and all species (Fig. 2, Supplementary Table 1). In *P. falciparum*,

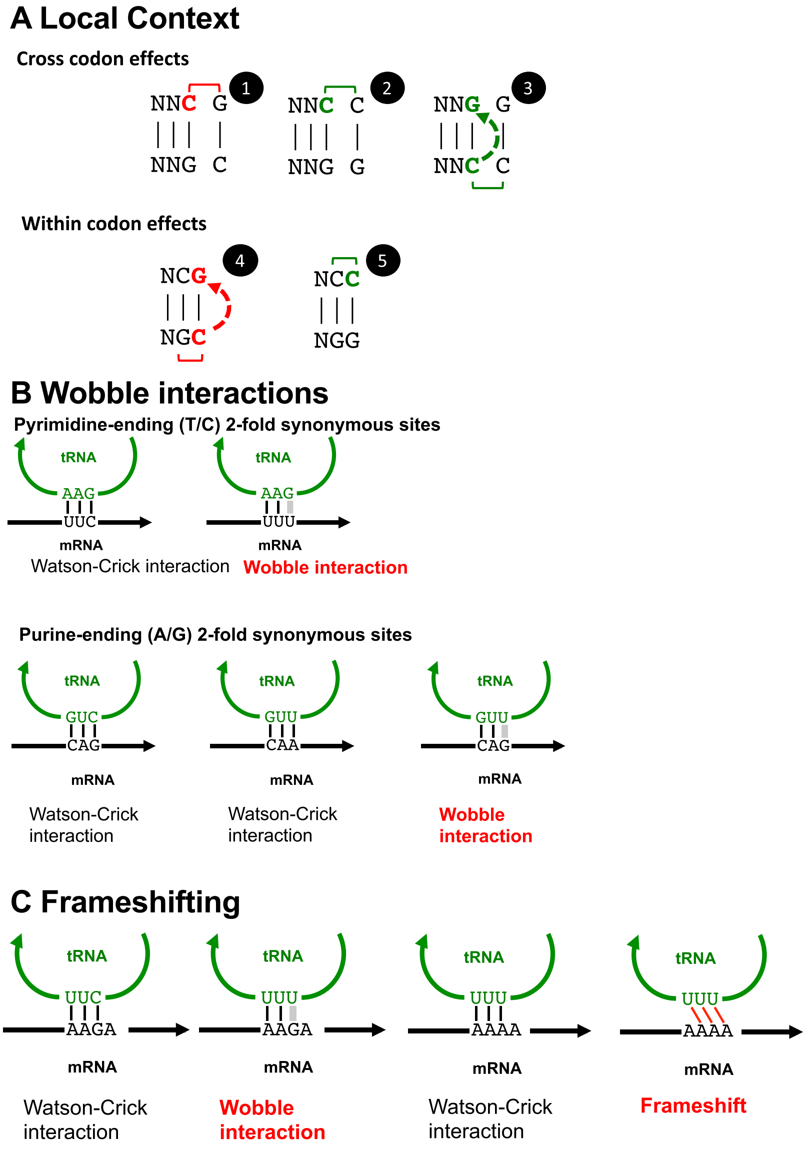

**Figure 1.** Proposed mechanisms of codon bias described. (**A**) Contextual effects. Across codons (free from translational selection) cytosines on the positive strand followed by an $N_1$ guanine (CpG) are suppressed (1) while cytosines followed by an $N_1$ cytosine (CpC) are protected (2). Guanines followed by an $N_1$ guanine are also more common, possibly reflecting protection of the complementary negative strand cytosine by its 5′ cytosine (3). 4-fold synonymous codons containing C2 (alanine, proline, threonine) show lower proportions of guanine in the 3rd position than glycine and valine, consistent with CpG suppression on the negative strand (4). These C2-containing amino acids also show higher proportions of C3 consistent with protection by the 5′ cytosine (5) on the positive stand. Key: Cytosines within CpG pairs are shown in red while cytosines within CpC pairs are shown in green; dotted arrows indicate effects deriving from the negative strand. (**B**) Avoidance of wobble tRNA-codon interactions in highly expressed genes. At pyrimidine-ending (T/C) 2-fold synonymous sites (phenylalanine shown here) the only tRNA available has G opposite the 3rd position of the codon, which binds NNC codons via Watson-Crick binding but NNU codons via G:U wobble, explaining the reduction in NNU codons in high expression genes. At purine-ending (A/G) 2-fold synonymous sites (glutamine shown here) both codons have cognate tRNAs but the NNG codon can also interact with the tRNA that has U opposite the 3rd position of the codon (U:G wobble), explaining the reduction in NNG codons in high expression genes. (**C**) Competing forces at lysine codons in highly expressed genes. The AAA codon should be favoured because it can only bind its cognate tRNA (avoiding wobble interactions), while the AAG codon can bind a near-cognate tRNA via wobble interaction. However if the next codon begins with an adenine nucleotide, the AAA codon risks frameshifting.

T3:C3 ratio is 3.8 for $N_1$C vs. 6.8 for $N_1$D (D = 'not C'). This suggests that there is protection of the positive strand C3 by the neighbouring cytosine nucleotide at $N_1$. T3:C3 is also relatively low with $N_1$A, perhaps because TA is suppressed.

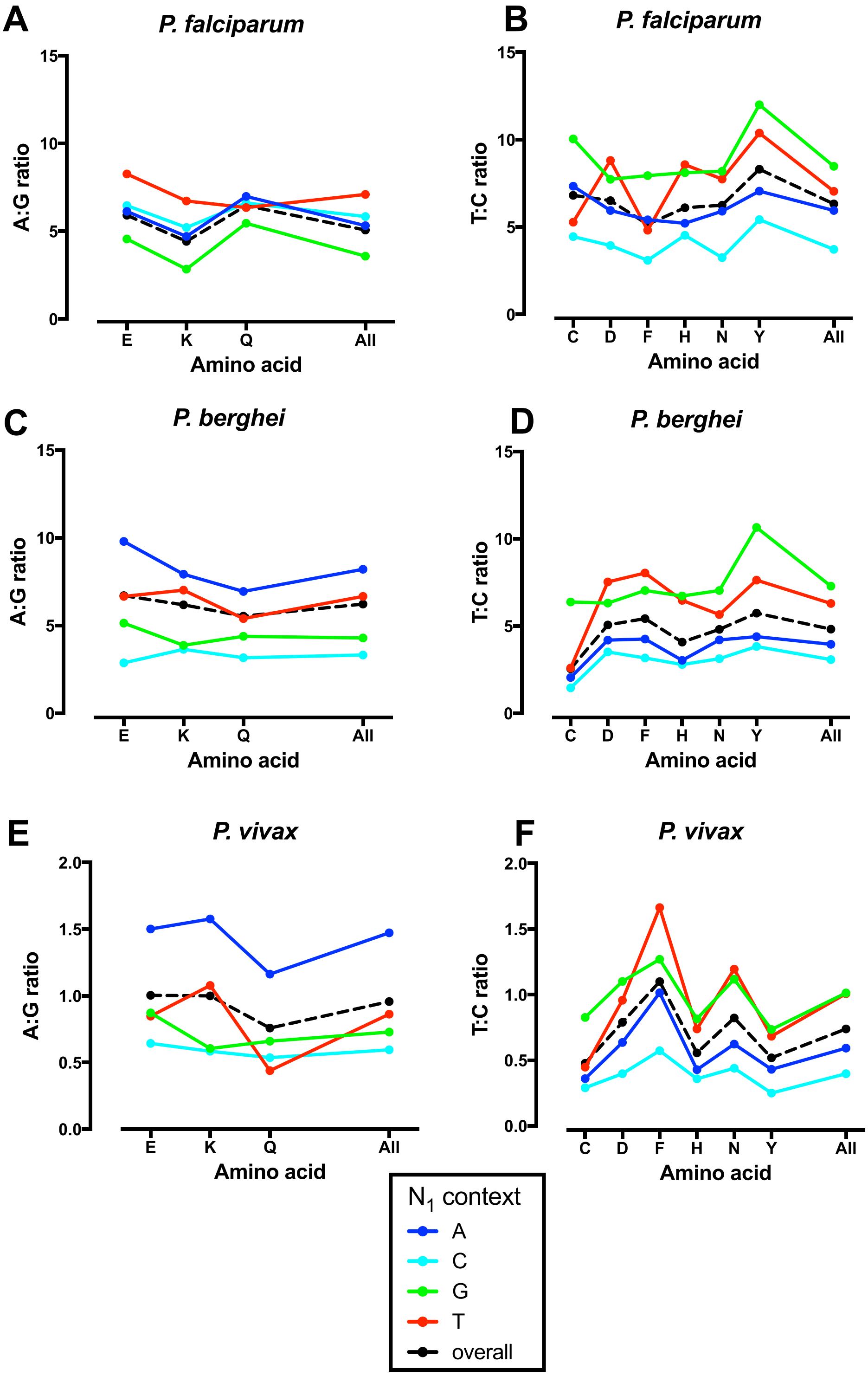

**Figure 2.** Effect of $N_1$ context at 2-fold synonymous sites. A3:G3 and T3:C3 ratios at 2-fold synonymous sites are shown for *P. falciparum* (**A**,**B**), *P. berghei* (**C**,**D**) and *P. vivax* (**E**,**F**).

$N_1$ context is also associated with an altered A3:G3 ratio (Fig. 2, Supplementary Table 1), although the effects are less consistent than with the T3:C3 ratio. In *P. falciparum* $N_1G$ is associated with the lowest A3:G3 ratio for all three amino acids ($N_1G = 3.6$ vs. 5.7 with $N_1H$ overall), a potential explanation being protection of the cytosine at position 3 on the negative strand by its neighbouring 5′ cytosine. For *P. berghei* and *P. vivax* $N_1G$ is also associated with a low A3:G3 ratio (Supplementary Table 1), although A3:G3 is even lower (overall) with $N_1C$, perhaps because AC is suppressed.

*4-fold sites (5 amino acids).* 4-fold synonymous sites were also examined to see if the associations seen at 2-fold sites are present at 4-fold sites. In the absence of simple ratios we assessed associations between $N_1$ context and the percentage of cytosines and guanines in the third position (C3% and G3%)(Supplementary Fig. 1, Supplementary Table 2).

As seen at 2-fold sites, cytosines are more rare with $N_1G$ for all species (6.2% vs. 9.2% with $N_1H$ for *P. falciparum*), consistent with CpG dinucleotides being reduced. Cytosines are also more common with $N_1C$ (C3% 10.2% vs. 8.3% for $N_1D$) again suggesting protection of C3 on the positive strand by the 3′ $N_1C$.

G3% is relatively less influenced by context at 4-fold sites. As at 2-fold sites, G3 is more common with $N_1G$ in *P. falciparum* (G3% 10.6% vs. 8.8% for $N_1H$), consistent with protection of C3 on the negative strand by the neighbouring (5′) cytosine on the negative strand of the next codon. Smaller magnitude but highly significant effects in the same direction are also seen in *P. berghei* and *P. vivax*.

*Within codon effects (4-fold sites, 5 amino acids).* The influence of $N_1$ context within codons can be assessed by examining whether within 4-fold synonymous sites the presence of a cytosine at position 2 (C2; alanine, proline, threonine) is associated with nucleotide usage at the third position compared to codons without C2 (termed D2; glycine and valine). Two associations observed in the cross-codon data are also visible within codons of *P. falciparum* (Supplementary Fig. 2, Supplementary Table 3). C2 is associated with a low G3% (7.2% vs. 11.5% with D2), consistent with vulnerability of negative strand cytosines in the third position. C2 is also associated with a high C3% (11.0% with C2 vs. 5.6% with D2), consistent with protection of C3 on the positive strand by the 5′ C2 cytosine.

There are similar findings with *P. berghei* (G3% 6.9% with C2 vs. 14.0% with D2; C3% 10.7% with C2 vs. 7.8% with D2) and *P. vivax* (G3% 28.4% with C2 vs. 38.5% with D2; C3% 34.8% with C2 vs. 23.6% with D2) (Supplementary Fig. 2C–F).

**Expression-related codon bias.** We next examined the relationship between expression and codon bias in the three malaria parasite species, focussing for clarity on 2-fold synonymous sites where nucleotide bias can be expressed as a simple ratio. Based on previous work[28], level of expression was grouped into three strata with the top stratum containing the top 3% of genes in terms of peak asexual phase expression. Intense expression in *P. falciparum* is generally associated with higher A3:G3 ratios (5.1 vs. 6.3 in the lowest and highest bands respectively), but lower T:C ratios (6.6 vs. 3.5) (Fig. 3A,B, Supplementary Tables 4 and 5). The fall in T3:C3 ratio with increasing expression is clear across all six relevant amino acids in *P. falciparum*, but the increase in A3:G3 ratio at higher levels of expression is not consistent, with the ratio approximately doubling for glutamatate and glutamine (producing a codon bias of more than 10-fold in the highest expression stratum) but falling significantly within lysine codons (4.4 to 4.1). There are similar findings using the *P. falciparum* microarray expression dataset (Supplementary Tables 4 and 5).

In *P. berghei*, there is a broadly similar picture, with consistent falls in T3:C3 ratio across all amino acids and similarly mixed findings at A3:G3 codons, with the fall in A3:G3 at lysine codons even more pronounced (Fig. 3C,D, Supplementary Tables 4 and 5). The preponderance of lysines compared to the other two amino acids means that in *P. berghei* the overall A3:G3 ratio at 2-fold A/G sites falls by around 12% in highly expressed genes.

For *P. vivax*, T3:C3 ratio only falls for some amino acids and the overall effect is of borderline significance. High level expression is associated with more significant changes in A3:G3 ratio, which rises in glutamate and glutamine codons but clearly falls at lysine codons.

In order to produce overall measures of selection for each organism, we calculated the strength of selection co-efficient *S* for the nine 2-fold synonymous codons, according to the approach of dos Reis and Wernisch[35], comparing the proportion of the codon selected in the top expression band (III) with its proportion in the lowest expression band (I). *S* is the natural log of the odds ratio of the relative codon frequencies in highly expressed compared to reference genes. For *P. falciparum* values between 0.09 and 1.07 are obtained for the nine codons (Table 2), with broad correlation (Spearman correlation across nine amino acids = 0.67, p = 0.059) between our results and those previously reported for *P. falciparum*[35] based on assumed expression levels of *P. falciparum* genes inferred by means of orthologous relationships (*S. cerevisiae* vs. *P. falciparum*).

For *P. berghei* the selection coefficients apply consistently in the same direction, with lower *S* values for most amino acids except lysine (see above). In *P. vivax*, *S* values are considerably smaller, varying from 0.092 to 0.304. Averaging across all amino acids (i.e. allowing for their relative frequencies in the entire coding sequence) we obtained averaged *S* values of 0.537 for *P. falciparum*, 0.422 for *P. berghei* and 0.199 for *P. vivax*.

**Stop codons and expression.** The first observation is that underlying AT-bias clearly influences stop codon usage with TAA codons making up around two thirds of *P. falciparum* and *P. berghei* stop codons but only one fifth of *P. vivax*, consistent with the overall AT-richness in the three species.

In *P. falciparum* the proportion of TAA stop codons increases from 67% to 94% in the top expression band while TAG and TGA codons fall significantly (Fig. 4, Supplementary Table 6). Tetranucleotides (the 'traditional' stop codon plus the following nucleotide) are also considered to be important in terms of efficiency of translational termination[37]. The TAAA tetranucleotide rises as a proportion of TAAN tetranucleotides (56% in the bottom expression band to 71% in the top) while TAAC and TAAT fall significantly and TAAG tetranucleotides show no significant change in proportion. Virtually identical associations were found when using ribosome profiling[24] to quantify expression during the asexual lifecyle (Supplementary Table 6).

Highly analogous changes are seen with *P. berghei*, with TAA stop codons rising from 64% to 89% across expression bands; again, the TAAA tetranucleotide rises (as a proportion of TAA), from 44% to 61% while TAAC and TAAT undergo significant falls.

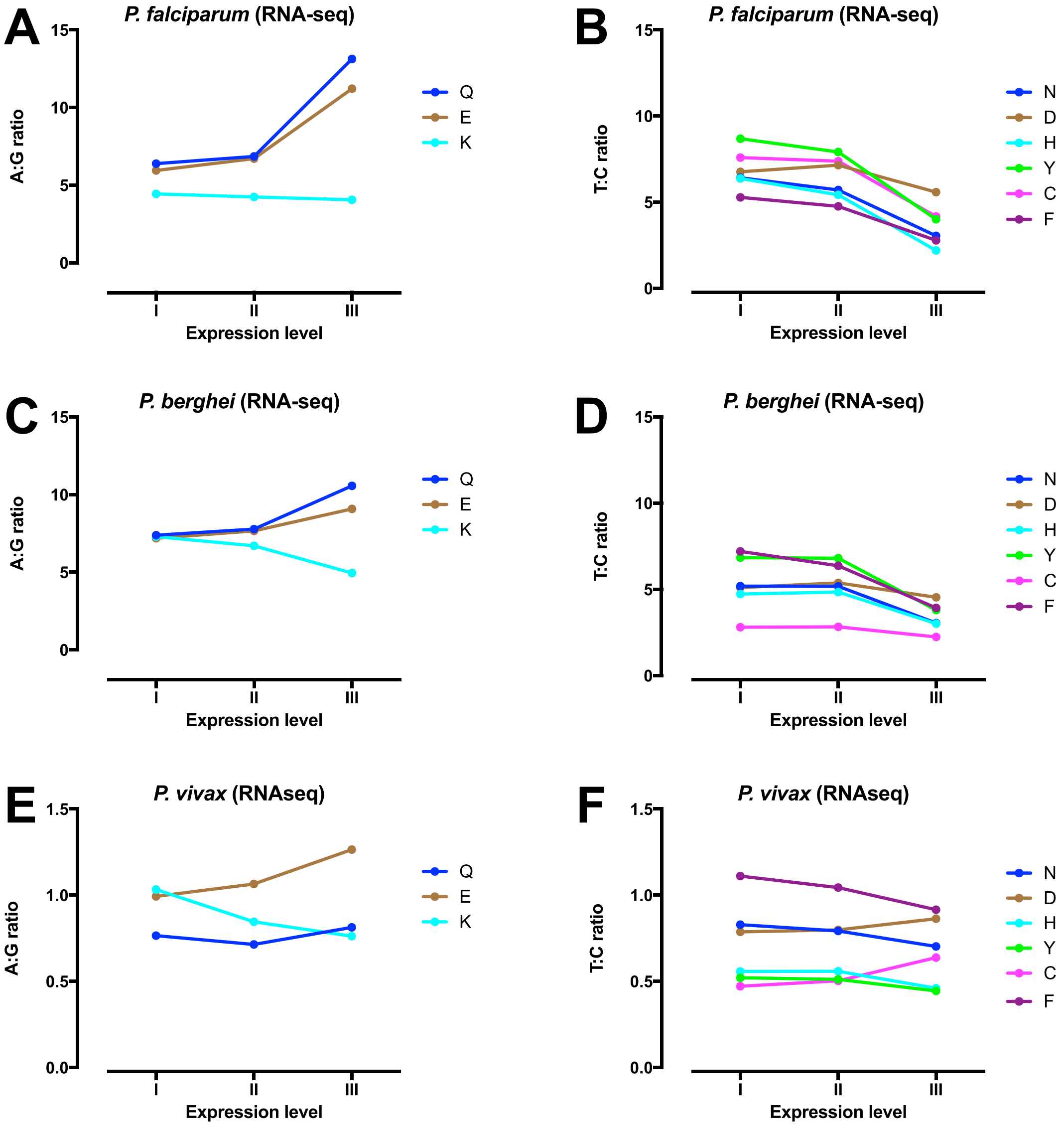

**Figure 3.** Effect of asexual gene expression at 2-fold synonymous sites. A3:G3 and T3:C3 ratios at 2-fold synonymous sites are shown for each of three expression levels for *P. falciparum* (**A**,**B**), *P. berghei* (**C**,**D**) and *P. vivax* (**E**,**F**). Results are presented for each individual amino acid.

| Amino acid | Selected codon in *P. falciparum* | *P. falciparum* | | *P. berghei* | *P. vivax* |
|---|---|---|---|---|---|
| | | This study (RNA-seq data) | Dos Reis & Wernisch (inferred by orthology) | This study (RNA-seq data) | This study (RNA-seq data) |
| E | GAA | 0.635 | 0.600 | 0.233 | 0.242 |
| K | AAG | 0.090 | 0.380 | 0.388 | 0.304 |
| Q | CAA | 0.720 | 0.420 | 0.358 | 0.061 |
| C | TGC | 0.598 | 1.100 | 0.225 | 0.301 (TGT) |
| D | GAC | 0.192 | 0.290 | 0.118 | 0.092 (GAT) |
| F | TTC | 0.641 | 0.530 | 0.610 | 0.194 |
| H | CAC | 1.067 | 1.030 | 0.450 | 0.192 |
| N | AAC | 0.744 | 1.350 | 0.534 | 0.165 |
| Y | TAC | 0.774 | 1.230 | 0.587 | 0.158 |
| Averaged *S* | | 0.537 | | 0.422 | 0.199 |

**Table 2.** Calculations of the strength of codon selection coefficient *S* for the nine 2-fold synonymous codons for *P. falciparum* and *P. vivax*, along with analogous values obtained by dos Reis & Wernisch[35]. For two amino acids in *P. vivax* the direction of effect is opposite to that in *P. falciparum*.

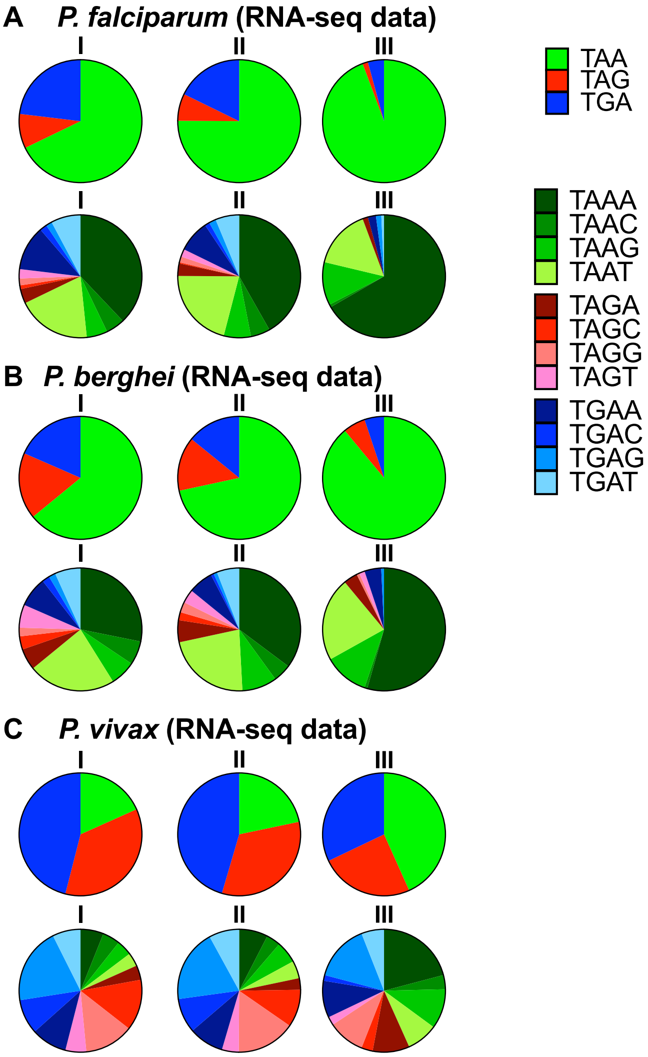

**Figure 4.** Effect of asexual gene expression and stop codon / tetranucleotide usage. Results are shown for each of three expression levels for *P. falciparum* (**A**), *P. berghei* (**B**) and *P. vivax* (**C**).

Changes in stop codon patterns of a similar direction, but smaller magnitude, are observed in *P. vivax* with TAA rising from 18% to 43% of stop codons in genes with high expression while TAG falls (Fig. 4, Supplementary Table 6). Within TAA tetranucleotides, there are again significant changes in TAAA proportion (rising with increased expression) and TAAC (falling) while TAAG and TAAT tetranucleotides are not significantly different.

### Combining intrinsic (AT-richness), contextual and expression-related biases.

Given the documented influence of overall AT-richness, local context and expression on codon bias in *P. falciparum*, we finally studied how these factors act in combination, in order to observe how overall AT-bias becomes more or less intense according to the prevailing set of conditions. For clarity of interpretation we again focused on changes in codon ratios at 2-fold synonymous sites (Figs 5 and 6).

The combination of underlying AT-richness and a high level of expression in *P. falciparum* are predicted to produce very high A3:G3 ratios; in glutamine and glutamate codons in highly expressed genes A3:G3 ratio is greater than 10, and for glutamine followed by $N_1C$ it increases to 38.8. As noted above, lysine codons behave differently; importantly the distinctive overall fall in A3:G3 ratio associated with high expression at lysine codons (see above) appears essentially attributable to codons where the following codon starts with an adenine ($N_1A$, which is present after 52% of all *P. falciparum* lysine codons). In this context A3:G3 falls from 4.8 in the lowest expression stratum to 3.0 in the highest. In other contexts ($N_1B$) A3:G3 in lysine codons rises with expression, in line with glutamine and glutamate. In *P. berghei* also, $N_1$ context clearly interacts with expression-related effects, with the $N_1A$ context again explaining much of the distinctive fall in A3:G3 ratio in lysines of highly expressed genes (with $N_1A$, A3:G3 falls from 10.8 in the lowest expression stratum to 5.2 in the highest).

The combined effects of $N_1$ context and the influence of expression on T3:C3 ratio in *P. falciparum* are also generally predictable from the individual effects (Fig. 6). The lowering of T3:C3 ratio in intensely expressed genes is countered in the $N_1G$ context (so that T3:C3 remains above 3 in all cases). However with $N_1C$ and intense expression (factors independently associated with preservation of C3), the two forces act in the same direction to

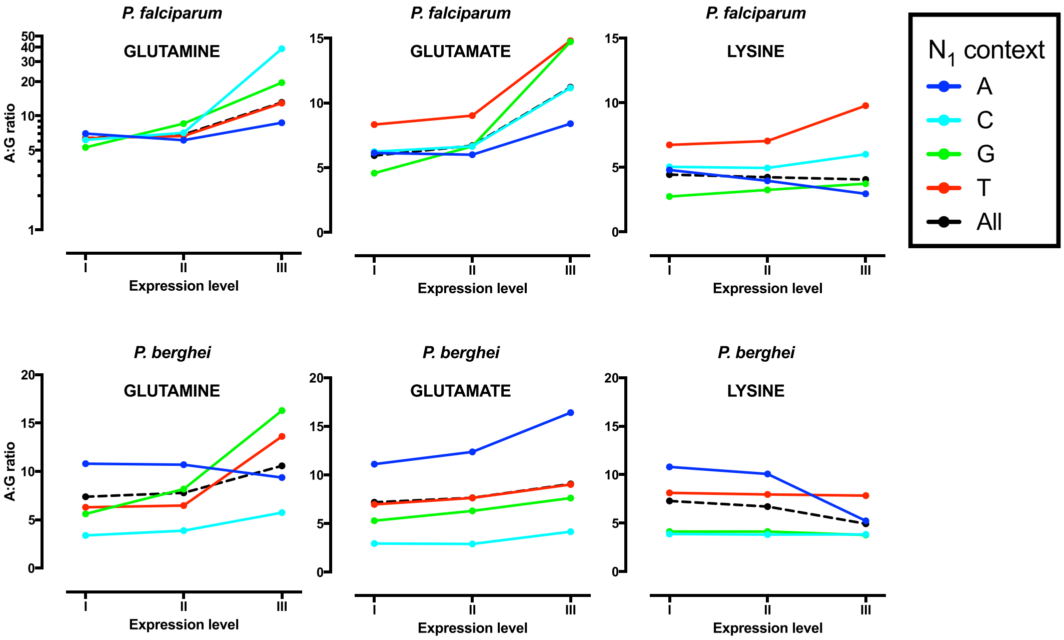

**Figure 5.** Contextual ($N_1$) and expression-related effects on A3:G3 for 2-fold synonymous sites in *P. falciparum* and *P. berghei*. For glutamine in *P. falciparum* the A:G ratio is shown on a log-scale.

lower T3:C3 ratio, producing an overall T3:C3 ratio across all codons of only 1.9. This is most extreme in phenylalanine codons followed by $N_1$C, which in highly expressed genes have a T3:C3 ratio of only 1.22. Similar patterns are observed with *P. berghei*, with the T:C ratio in an $N_1$C context reduced to 2.0 (across all six amino acids) and the AT-bias at cysteines followed by $N_1$C disappearing entirely in highly expressed genes (T3:C3 = 0.87).

## Discussion

Genomewide association studies focus on non-synonymous changes, with synonymous mutations typically not considered as likely causes of phenotypic change[38]. In previous work we showed that many non-synonymous polymorphisms in the non-conserved sections of *P. falciparum* coding sequence are likely to be non-adaptive in nature[39], emphasising the importance of neutral evolution in *P. falciparum* and suggesting how taking this into account might be of benefit in assessing drug-resistance mutations in population genetic studies of *P. falciparum*[40]. In contrast, this work, and other key studies which preceded it[27,28,41], indicate that synonymous mutations in highly expressed genes (including their stop sequences) are not simply neutral.

This study builds on previous observations on codon bias in malaria parasites, revealing how codon bias is influenced by several complex 'layers' of mutational and selective forces that differ between malaria species. The most obvious (and 'deepest') layer of bias is the overall AT-richness of the genome, with extreme AT-richness found in both the Laverania group (exemplified by *P. falciparum*) and rodent parasites (e.g. *P. berghei*). In *P. falciparum* this property is maintained by a mutational force, with both population genetic studies[42] and studies of mutation in continuous cultures[43] indicating that the genome is broadly in equilibrium, although this bias might originally have been selected.

A second clear 'layer' of influence on codon bias is local DNA context. Irrespective of the underlying AT-bias of each genome, cytosines are the most vulnerable nucleotide[44] and the probability of mutation (presumably via deamination) is context-dependent. There is a high T3:C3 ratio at synonymous positions followed by a 3' guanine nucleotide, consistent with CpG dinucleotide suppression, a finding previously reported in focused[29,31] and genome-wide studies[32]. In organisms with DNA methylation, CpG suppression is frequently attributed to enhanced spontaneous deamination of methylated cytosines (5-mC) to thymine, which results in a T:G mismatch that can then be repaired, or replicated, to give a C:G to T:A substitution. However there is conflicting evidence as to whether DNA methylation occurs in malaria parasites[30,45,46] and other Apicomplexans[47]. If the CpG suppression is not due to methylation, it may simply represent a methylation-independent mutational hotspot. CpG suppression clearly occurs in yeast in the absence of DNA methylation[48].

Mutational bias in the absence of methylation would also explain other distinct contextual effects that we observe. In contrast to the vulnerability of cytosine preceding guanines, cytosines appear to be protected when another cytosine is neighbouring, consistent with the previous finding that in coding regions of a number of *P. falciparum* genes, CC dinucleotides are more common than expected[29]. The effect is particularly strong when a cytosine is found on the 3' side. In all species and for all amino acids, the T3:C3 ratio at 2-fold synonymous codons is reduced with the presence of an $N_1$ cytosine; there is additional supportive evidence from 4-fold codons. We also note that a 5′ cytosine protects cytosines, with higher levels of G3 with $N_1$G, consistent with protection of C3 on the negative strand by the neighbouring C on the negative strand of the next codon (5′). This is strengthened further by within-codon data, showing that the 4-fold amino acids with C2 (alanine, proline and threonine) have

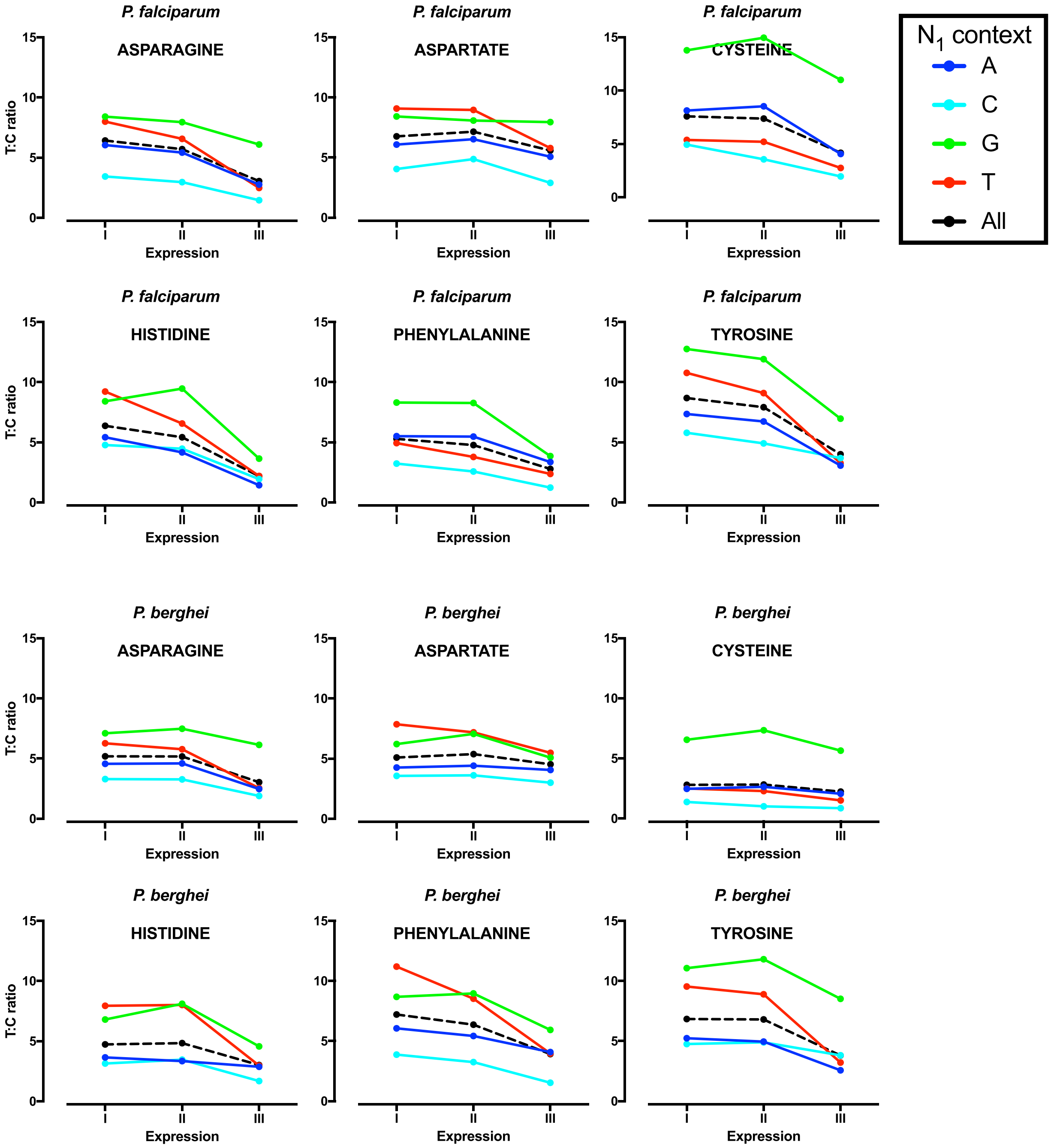

**Figure 6.** Contextual ($N_1$) and expression-related effects on T3:C3 for 2-fold synonymous sites in *P. falciparum* and *P. berghei*.

a signficantly higher proportion of C3 than others (glycine and valine). Additional effects, such as T3:C3 being relatively low with $N_1$A, and A3:G3 being lowest with $N_1$G in *P. berghei* and *P. vivax*, are consistent with previous work on dinucleotides[29] and indicate that a variety of local contextual forces influence synonymous positions. These contextual findings are important when interpreting changes in codon bias in genes with a high level of expression in the asexual cycle (see below).

Our study of the influence of expression on codon bias[27,28,35] extends previous work by including more recent RNA-seq data pertaining to three malaria species. As previously reported for *P. falciparum*, high level expression in *P. falciparum* and *P. berghei* is associated with a broad increase in A3:G3 ratio and reduction in T3:C3 ratio at 2-fold synonymous sites[27,28]. Assuming that selection drives these effects, what are the likely mechanisms? Musto *et al.*, who first described distinctive patterns of codon usage according to level of gene expression[27], suggested that the selective force could be optimisation of translation according to available tRNAs, since for pyrimidine (C or T-ending codons) the only existing isoacceptor tRNA binds perfectly with the significantly incremented triplet[28] (Fig. 1B). Recent work by Chan *et al.* during the conduct of our study[41] explores this further, in particular at asparagine codons where AAT codons can only be decoded via a near-cognate tRNA predicted to bind only via a 'wobble' interaction[49]. Reading of codons by wobble binding in metazoans is associated with slower translation because of ribosomal stalling[11]. By studying asparigine homorepeats, Chan *et al.* were able to show directly that wobble pairings also reduce translation efficiency in *Plasmodium falciparum*[41]. Hence the fall in T3:C3 ratio in high expression genes (also reported by Chan *et al.* to be greater in *P. falciparum* than *P. vivax*) is readily explained by selection against such codons.

For purine (A- or G-ending) codons, malaria parasites have tRNA isoacceptors cognate for both possible codons. The 'wobble' hypothesis might again be relevant, since any NNG codons can not only bind their cognate C-starting tRNA anticodon, but also the near-cognate U-starting tRNA anticodon (see Fig. 1B), via wobble interaction. Increased use of NNA codons that bind only their cognate tRNA would prevent such pairing (with its attendant reduction in translational efficiency), providing a selective hypothesis for the higher A3:G3 ratio. Avoidance of wobble interactions might also contribute to the higher A3:G3 ratios in the highly expressed genes of certain other organisms; for example in *S. cerevisiae*, for Gln there are 9 UUG tRNAs (cognate to CAA codons) and only one tRNA for CUG (cognate to CAG codons)[35]; selection of CAA, which can only bind the UUG tRNAs, would reduce the number of wobble interactions.

In *P. vivax* the same overall changes take place in high expression genes, although they are of lesser magnitude and less consistent across the various amino acids in each group. This fits with data from Chan *et al.* which report minimal reduction in use of wobble codons in highly expressed *P. vivax* genes[41].

An interesting exception to the broad patterns of codon bias associated with high gene expression is the clear fall in A3:G3 ratio in lysine codons; this is evident for all three malaria species studied, being most clearly visible in *P. berghei*. Chan *et al.* also found that lysine codons have distinctive properties (lower use of A-ending codons) in ribosomal proteins, considered as reference genes for describing optimal codon usage. How can this pattern be explained? Local DNA context appears to be critical, since in both *P. falciparum* and *P. berghei* the presence of an adenine at the start of the next codon ($N_1A$) is associated with substantially lower A3:G3 ratio in high expression genes, while the other three $N_1$ possibilities produce more typical increases in A3:G3 ratio. A logical explanation for this is that in highly expressed genes, where there is selection for the AAA codon (binding the single UUU isoacceptor hence avoiding wobble binding), there is also competing selection against the AAAA tetranucleotide, since after binding of the UUU isoacceptor tRNA there can be + 1 frameshifting (Fig. 1C). Around half of lysine codons are followed by an A at the start of the next codon so this is a significant force. Frameshifting at lysine codons has been described in bacteria[8], and frequent translational errors of this form would clearly produce devastating consequences in highly expressed genes[50], with waste of energy to generate and then degrade high levels of non-functional peptide chains; misfolded proteins also risk cell toxicity[51].

We also studied stop codon usage, and how this changes in high expression genes, with the idea that this might also provide an insight into generalised forces acting on DNA, as well as the distinctive biology of translational termination[5]. Forces acting upstream or downstream of translation should apply equally to stop and sense codons, whereas if selection on codon bias is principally associated with tRNA frequencies or interactions, the stop codons may behave differently in highly expressed genes. The first observation was that TAA codons make up around two thirds of *P. falciparum* and *P. berghei* stop codons but only one fifth of *P. vivax*, consistent with the overall AT-richness in the three species. In all three species of malaria parasites, highly expressed genes show increasing use of the TAA codon, with the bias in *P. falciparum* and *P. berghei* particularly striking (the proportion of TAA codons rising to 94 and 89% respectively). This is of interest as across a wide range of species such relationships between expression and stop codon usage have been hard to discern[52], and suggests that the (presumably) selective forces are relatively strong in these malaria species. TAA is also favoured in highly expressed genes in humans[53]. The simplest mechanistic explanation for increasing TAA relates to efficiency[54], as TAA is the 'universal' stop codon and binds to either of the relevant release factors, a phenomenon well documented in bacteria[55].

Translational termination is further optimised by the flanking base at the downstream position (+1) so that the 'stop tetranucleotide' can be considered to signal the termination of protein synthesis in eukaryotes[37]. In highly expressed genes of all three malaria species, TAAA tetranucleotides occupy a significantly higher proportion within the overall set of TAA trinucleotide stop codons, while TAAC (and TAAT in *P. falciparum* and *P. berghei*) become significantly less common; TAAG is not significantly changed. These findings match those obtained in mammalian systems where the order for termination efficiency of the base at +1 position was found to be $A \approx G \gg C \approx U$ (independently of the stop codon), UAAA being the most efficient four-base combination[56].

These expression-related effects are likely to promote fitness by improving efficiency of translation and hence producing more competitive parasites. Their strength differs according to species; we find that selection parameter *S* (averaged across amino acids) is largest in *P. falciparum*, slightly smaller in *P. berghei* and substantially smaller again in *P. vivax*, with values generally lower than for other typical eukaryotic genomes[35]. We note that this ranking matches the relative overall AT-richness of the respective genomes. This correlation, which could potentially be explored across a wider range of malaria species, suggests that AT-richness itself might also be a fitness adaptation promoting replication 'efficiency'; for example the genomes of bacteria that rely on their host for survival tend to be AT-rich given the increased energy cost in GTP and CTP synthesis[57]. Hence overall AT-richness, codons that avoid tRNA wobble interactions and frameshifts, and specific stop sequences may all provide small competitive advantages in terms of speed and efficiency of asexual replication.

Why do these processes operate at different levels in the different malaria species? The answer might relate to lifecycle biology, which has distinct characteristics according to parasite species. One prominent example is the relatively early gametocytogenesis of human *P. vivax* infection, in which transcriptional control promotes early conversion to the sexual stage; this might therefore reduce selection for maximal asexual cycle efficiency[58]. Alternatively, the effective population size of the organism over evolutionary time might be responsible for the differences in synonymous codon usage. The probability that a mutation will change in frequency depends on the product of the effective population size and the true selection coefficient (*s*)[59] (the parameter *S* which we describe above is actually a 'confounded' selection co-efficient that is influenced by effective population size). There are several examples of codon usage bias being determined by effective population size in other eukaryotes[60], with shifts in codon bias tending to occur when organisms have a large effective population size that facilitates the selection of mild-effect polymorphisms[35]. By contrast, a prolonged population bottleneck may result in reduced selection and low levels of codon bias even in highly expressed genes. There have been substantial advances in our understanding of the evolutionary histories of malaria parasites[61] in recent years, but a description of how effective population size has varied in each species and its ancestors remains some way off. Interestingly, a recent study based on MalariaGEN Community Project data suggests a substantially higher preference for AT nucleotides compared to GC nucleotides at synonymous single nucleotide polymorphism sites in *P. vivax*[62], a process that in principle should produce a more AT-rich genome over time. This is consistent with the finding that *P. vivax* has recently undergone a population expansion relative to *P. falciparum*[58,63].

**Limitations.** Our study was of descriptive design and hence did not meet all criteria allowing distinction between association and causality, particularly relevant if selective forces are being invoked[64]. It is therefore appropriate to consider whether there might be mutational explanations for phenomena where we propose selective hypotheses. For example, bacterial expression can influence codon usage because the DNA of highly expressed genes spends a relatively greater proportion of its time in single-stranded form, making it more prone to mutational forces[65]. Strand bias in cytosine deamination has been postulated to play a major role in genome evolution[66] with cytosines less common in the non-transcribed strand of highly expressed genes; however this is the opposite of the pattern we observe in *P. falciparum*. The similarity between the patterns of codon bias we found in highly expressed genes, and analogous studies in other higher eukaryotes[35], also supports selective explanations.

In terms of the influence of local DNA context on codon bias, we assume the opposite i.e. that context introduces mutational 'hotspots' (CpG) or 'coldspots' (CpC). One factor supporting this assumption is that effects appear consistent across multiple amino acids, and species, and follow simple rules at the DNA level. Again, direct evidence that mutational force mediates these contextual effects remains lacking. There appears to be no statistically significant enrichment of CpG or TpC dinucleotides at de novo C → T transitions in cultured *P. falciparum* lines[43], although the power of that study was probably not sufficient to discern an additional contextual effect beyond the underlying AT-bias. Large scale population genetic data might shed more light on this area[67].

Our study necessarily focused on certain categories of variation in DNA context or expression. We looked at a series of neighbouring base contextual effects, but the constraints of the genetic code meant that certain contexts were not examined comprehensively, including preceding (−1) bases, and more distant bases. For studies of codon bias in highly expressed genes we focused on 2-fold codons as interpretation is relatively simpler than in 4-fold sites where complexity of competing forces is likely to make interpretation challenging. Levels of protein expression may also be influenced by other adaptive changes, for example in RNA stability and tRNA modification[68], factors not explored in this study.

Our primary analyses focused upon the transcriptome as assessed by RNA-seq methodology, given the availability of comparable datasets for all three studied species. In the case of *P. falciparum* we were able also to look at the association between codon bias and level of mRNA translation using ribosome profiling data[24], comparing the results to data from RNA-seq.[20]; for convenience we focused on stop codons. Given the tight coupling between transcription and translation, particularly for the core genome[24], it was no surprise to see virtually identical results across RNA-seq and ribosome profiling data, confirming that TAA stop codon (and the TAAA tetranucleotide) is enriched in genes that are highly transcribed and translated. Our method to rank genes by level of expression was relatively simple and based on previous work, taking the maximum level expression at any point in the asexual cycle as a particular gene's value for ranking purposes[28]. In their recently published work, Chan *et al.* examined the top 5% expressed genes during the progression of the asexual cycle in a highly analogous approach to ours[41].

## Conclusions

The various forms of codon bias in malaria parasites result from a complex set of forces that produce a variegated bias within and between genes. Local DNA context exerts clear effects, with CpG dinucleotides associated with cytosine depletion and CpC dinucleotides relative cytosine protection. Genes that are expressed intensely during asexual stages are affected by additional forces that are consistent with promotion of translational efficiency through avoidance of wobble interactions during tRNA binding, reduction of frameshifting and optimisation of termination. These expression-related biases are more substantial in the AT-rich genomes of *P. falciparum* and *P. berghei* than in the less AT-rich genome of *P. vivax*, suggesting that underlying properties of the different species, such as the competitive advantage offered by rapid replication or overall population size, influence both AT-richness and codon bias.

## References

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* **8**, r49–r62 (1980).
2. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. *Proc Natl Acad Sci USA* **76**, 1697–1701 (1979).
3. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**, 389–409 (1981).
4. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9**, r43–74 (1981).
5. Trotta, E. Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Res* **41**, 9382–9395, https://doi.org/10.1093/nar/gkt740 (2013).

6. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**, 98–108, https://doi.org/10.1038/nrg1770 (2006).
7. Rodriguez, O., Singh, B. K., Severson, D. W. & Behura, S. K. Translational selection of genes coding for perfectly conserved proteins among three mosquito vectors. *Infect Genet Evol* **12**, 1535–1542, https://doi.org/10.1016/j.meegid.2012.06.005 (2012).
8. Berg, O. G. & Silva, P. J. Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Res* **25**, 1397–1404 (1997).
9. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32–42, doi:nrg2899 (2011).
10. Sharp, P. M., Emery, L. R. & Zeng, K. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci* **365**, 1203–1212, https://doi.org/10.1098/rstb.2009.0305 (2010).
11. Stadler, M. & Fire, A. Wobble base-pairing slows *in vivo* translation elongation in metazoans. *RNA* **17**, 2063–2073, https://doi.org/10.1261/rna.02890211 (2011).
12. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
13. Sueoka, N. Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. *Proc Natl Acad Sci USA* **47**, 1141–1149 (1961).
14. Rocha, E. P. & Feil, E. J. Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genet* **6**, e1001104, https://doi.org/10.1371/journal.pgen.1001104 (2010).
15. White, N. J. *et al*. Malaria. *Lancet* **383**, 723–735, https://doi.org/10.1016/S0140-6736(13)60024-0 (2014).
16. Gardner, M. J. *et al*. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511, https://doi.org/10.1038/nature01097 (2002).
17. Carlton, J. M. *et al*. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* **455**, 757–763, doi:nature07327 (2008).
18. Cheeseman, I. H. *et al*. A major genome region underlying artemisinin resistance in malaria. *Science* **336**, 79–82, doi:336/6077/79 (2012).
19. Bozdech, Z. *et al*. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol* **1**, E5, https://doi.org/10.1371/journal.pbio.0000005 (2003).
20. Otto, T. D. *et al*. New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. *Mol Microbiol* **76**, 12–24, doi:MMI7026 (2010).
21. Bozdech, Z. *et al*. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc Natl Acad Sci USA* **105**, 16290–16295, https://doi.org/10.1073/pnas.0807404105 (2008).
22. Zhu, L. *et al*. New insights into the Plasmodium vivax transcriptome using RNA-Seq. *Sci Rep* **6**, 20498, https://doi.org/10.1038/srep20498 (2016).
23. Otto, T. D. *et al*. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol* **12**, 86, https://doi.org/10.1186/s12915-014-0086-0 (2014).
24. Caro, F., Ahyong, V., Betegon, M. & DeRisi, J. L. Genome-wide regulatory dynamics of translation in the Plasmodium falciparum asexual blood stages. *Elife* **3**, https://doi.org/10.7554/eLife.04106 (2014).
25. Bastien, O. *et al*. Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference. *Gene* **336**, 163–173, https://doi.org/10.1016/j.gene.2004.04.029 (2004).
26. Carlton, J., Silva, J. & Hall, N. The genome of model malaria parasites, and comparative genomics. *Curr Issues Mol Biol* **7**, 23–37 (2005).
27. Musto, H., Romero, H., Zavala, A., Jabbari, K. & Bernardi, G. Synonymous codon choices in the extremely GC-poor genome of Plasmodium falciparum: compositional constraints and translational selection. *J Mol Evol* **49**, 27–35 (1999).
28. Peixoto, L., Fernandez, V. & Musto, H. The effect of expression levels on codon usage in Plasmodium falciparum. *Parasitology* **128**, 245–251 (2004).
29. Hyde, J. E. & Sims, P. F. Anomalous dinucleotide frequencies in both coding and non-coding regions from the genome of the human malaria parasite Plasmodium falciparum. *Gene* **61**, 177–187 (1987).
30. Pollack, Y., Kogan, N. & Golenser, J. Plasmodium falciparum: evidence for a DNA methylation pattern. *Exp Parasitol* **72**, 339–344 (1991). doi:0014-4894(91)90079-C.
31. Schorderet, D. F. & Gartler, S. M. Analysis of CpG suppression in methylated and nonmethylated species. *Proc Natl Acad Sci USA* **89**, 957–961 (1992).
32. Neafsey, D. E., Hartl, D. L. & Berriman, M. Evolution of noncoding and silent coding sites in the Plasmodium falciparum and Plasmodium reichenowi genomes. *Mol Biol Evol* **22**, 1621–1626, doi:msi154 (2005).
33. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**, 1499–1504 (1980).
34. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* **33**, 1141–1153, https://doi.org/10.1093/nar/gki242 (2005).
35. dos Reis, M. & Wernisch, L. Estimating translational selection in eukaryotic genomes. *Mol Biol Evol* **26**, 451–461, https://doi.org/10.1093/molbev/msn272 (2009).
36. Yadav, M. K. & Swati, D. Comparative genome analysis of six malarial parasites using codon usage bias based tools. *Bioinformation* **8**, 1230–1239, https://doi.org/10.6026/97320630081230 (2012).
37. Brown, C. M., Stockwell, P. A., Trotman, C. N. & Tate, W. P. Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Res* **18**, 6339–6345 (1990).
38. Hurst, L. D. Molecular genetics: The sound of silence. *Nature* **471**, 582–583, https://doi.org/10.1038/471582a (2011).
39. Gardner, K. B. *et al*. Protein-based signatures of functional evolution in Plasmodium falciparum. *BMC Evol Biol* **11**, 257, doi:1471-2148-11-257 (2011).
40. MalariaGEN *Plasmodium falciparum* Community Project. Genomic epidemiology of artemisinin resistant malaria. *Elife* **5** https://doi.org/10.7554/eLife.08714 (2016).
41. Chan, S., Ch'ng, J. H., Wahlgren, M. & Thutkawkorapin, J. Frequent GU wobble pairings reduce translation efficiency in Plasmodium falciparum. *Sci Rep* **7**, 723, https://doi.org/10.1038/s41598-017-00801-9 (2017).
42. Chang, H. H. *et al*. Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. *Mol Biol Evol* **29**, 3427–3439, https://doi.org/10.1093/molbev/mss161 (2012).
43. Hamilton, W. L. *et al*. Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Res* **45**, 1889–1901, https://doi.org/10.1093/nar/gkw1259 (2017).
44. Kreutzer, D. A. & Essigmann, J. M. Oxidized, deaminated cytosines are a source of C– > T transitions *in vivo*. *Proc Natl Acad Sci USA* **95**, 3578–3582 (1998).
45. Choi, S. W., Keyes, M. K. & Horrocks, P. LC/ESI-MS demonstrates the absence of 5-methyl-2′-deoxycytosine in Plasmodium falciparum genomic DNA. *Mol Biochem Parasitol* **150**, 350–352 doi:S0166-6851(06)00221-0 (2006).
46. Ponts, N. *et al*. Genome-wide mapping of DNA methylation in the human malaria parasite Plasmodium falciparum. *Cell Host Microbe* **14**, 696–706, https://doi.org/10.1016/j.chom.2013.11.007 (2013).
47. Gissot, M., Choi, S. W., Thompson, R. F., Greally, J. M. & Kim, K. Toxoplasma gondii and Cryptosporidium parvum lack detectable DNA cytosine methylation. *Eukaryot Cell* **7**, 537–540 doi:EC.00448-07 (2008).

48. Behringer, M. G. & Hall, D. W. Genome-Wide Estimates of Mutation Rates and Spectrum in Schizosaccharomyces pombe Indicate CpG Sites are Highly Mutagenic Despite the Absence of DNA Methylation. *G3 (Bethesda)* **6**, 149–160, https://doi.org/10.1534/g3.115.022129 (2015).
49. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol Rev* **56**, 229–264 (1992).
50. Urbonavicius, J., Qian, Q., Durand, J. M., Hagervall, T. G. & Bjork, G. R. Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J* **20**, 4863–4873, https://doi.org/10.1093/emboj/20.17.4863 (2001).
51. Huang, Y., Koonin, E. V., Lipman, D. J. & Przytycka, T. M. Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. *Nucleic Acids Res* **37**, 6799–6810, https://doi.org/10.1093/nar/gkp712 (2009).
52. Sun, J., Chen, M., Xu, J. & Luo, J. Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. *J Mol Evol* **61**, 437–444, https://doi.org/10.1007/s00239-004-0277-3 (2005).
53. Trotta, E. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC Genomics* **17**, 366, https://doi.org/10.1186/s12864-016-2692-4 (2016).
54. Manuvakhova, M., Keeling, K. & Bedwell, D. M. Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *RNA* **6**, 1044–1055 (2000).
55. Korkmaz, G., Holm, M., Wiens, T. & Sanyal, S. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem* **289**, 30334–30342, https://doi.org/10.1074/jbc.M114.606632 (2014).
56. McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J. & Tate, W. P. Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc Natl Acad Sci USA* **92**, 5431–5435 (1995).
57. Rocha, E. P. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet* **18**, 291–294, https://doi.org/10.1016/S0168-9525(02)02690-2 (2002).
58. Parobek, C. M. *et al*. Selective sweep suggests transcriptional regulation may underlie Plasmodium vivax resilience to malaria control measures in Cambodia. *Proc Natl Acad Sci USA* **113**, E8096–E8105, https://doi.org/10.1073/pnas.1608828113 (2016).
59. Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu Rev Genet* **42**, 287–299, https://doi.org/10.1146/annurev.genet.42.110807.091442 (2008).
60. Akashi, H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**, 927–935 (1994).
61. Loy, D. E. *et al*. Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *Int J Parasitol* **47**, 87–97, https://doi.org/10.1016/j.ijpara.2016.05.008 (2017).
62. Goel, P. & Singh, G. P. Divergent pattern of genomic variation in Plasmodium falciparum and P. vivax. *F1000 Research* **5**, 2763 (2016).
63. Jennison, C. *et al*. Plasmodium vivax populations are more genetically diverse and less structured than sympatric Plasmodium falciparum populations. *PLoS Negl Trop Dis* **9**, e0003634, https://doi.org/10.1371/journal.pntd.0003634 (2015).
64. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**, 640–649 (2002).
65. Francino, M. P., Chao, L., Riley, M. A. & Ochman, H. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272**, 107–109 (1996).
66. Beletskii, A. & Bhagwat, A. S. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proc Natl Acad Sci USA* **93**, 13919–13924 (1996).
67. Manske, M. *et al*. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**, 375–379, https://doi.org/10.1038/nature11174 (2012).
68. Ng, C. S. *et al*. tRNA epitranscriptomics and biased codon are linked to proteome expression in Plasmodium falciparum. *Mol Syst Biol* **14**, e8009, https://doi.org/10.15252/msb.20178009 (2018).

## Acknowledgements

The authors would like to thank Professor Nick White and Professor Peter Preiser for useful discussions.

## Author Contributions

Both authors designed the analyses. I.S. undertook coding and C.W. prepared the figures. Both authors wrote and reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at https://doi.org/10.1038/s41598-018-34404-9.

**Competing Interests:** The authors declare no competing interests.

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