Genetica (2021) 149:217–237
https://doi.org/10.1007/s10709-021-00126-6

ORIGINAL PAPER

Evolutionary pressures and codon bias in low complexity regions of plasmodia

Andrea Cappannini[1] · Sergio Forcelloni[3,4] · Andrea Giansanti[1,2]

Received: 4 September 2020 / Accepted: 30 June 2021

Abstract

The biological meaning of low complexity regions in the proteins of *Plasmodium* species is a topic of discussion in evolutionary biology. There is a debate between selectionists and neutralists, who either attribute or do not attribute an effect of low-complexity regions on the fitness of these parasites, respectively. In this work, we comparatively study 22 *Plasmodium* species to understand whether their low complexity regions undergo a neutral or, rather, a selective and species-dependent evolution. The focus is on the connection between the codon repertoire of the genetic coding sequences and the occurrence of low complexity regions in the corresponding proteins. The first part of the work concerns the correlation between the length of plasmodial proteins and their propensity at embedding low complexity regions. Relative synonymous codon usage, entropy, and other indicators reveal that the incidence of low complexity regions and their codon bias is species-specific and subject to selective evolutionary pressure. We also observed that protein length, a relaxed selective pressure, and a broad repertoire of codons in proteins, are strongly correlated with the occurrence of low complexity regions. Overall, it seems plausible that the codon bias of low-complexity regions contributes to functional innovation and codon bias enhancement of proteins on which *Plasmodium* species rest as successful evolutionary parasites.

Keywords Low complexity regions · Evolutionary pressures · Codon bias · Plasmodium · Malaria

Introduction

Malaria is caused by unicellular protozoans belonging to the genus *Plasmodium*, which comprises more than 200 species that parasitize a range of vertebrate hosts: reptiles, birds, and mammals. *Plasmodia* are successful parasites that have a large, dreadful economic and clinical impact on the human species. High genetic flexibility is thought to be the common trait of the *Plasmodium* species which allows them to develop resistance to antimalarial treatments and change host specificity (Sato 2021). It is then relevant to understand the common genetic mechanisms that confer to *Plasmodia* their adaptiveness. In this article, we present the results of a comparative study of low complexity regions (LCRs) in the proteins of *Plasmodia* and their biology. LCRs are amino acid sequences that contain repeats of single amino acids or short amino acid motifs (Toll-Riera et al. 2012). In *Plasmodium* species, low complexity regions have been associated with phenotypic plasticity and the evolution of host-pathogen interactions (Chaudhry et al. 2018). It is then interesting to correlate the occurrence and stability of LCRs, at the proteomic level, with the codon bias of the corresponding coding regions, at the genomic level. Two main mechanisms have been proposed to explain the origin and fixation of LCRs. The first, *replication slippage* (Gemayel et al. 2012; Saitou 2018), is a mutational process that occurs during DNA replication, involving denaturation and displacement of DNA strands with the consequent decoupling of complementary bases (Levinson and Gutman 1987). The other

✉ Andrea Cappannini
andreacappannini@gmail.com

Sergio Forcelloni
forcelloni@biochem.mpg.de

Andrea Giansanti
andrea.giansanti@roma1.infn.it

[1] Department of Physics, Sapienza, University of Rome, P.le A. Moro 5, 00185 Roma, Italy

[2] Istituto Nazionale di Fisica Nucleare, INFN, Roma1 section. 00185, Roma, Italy

[3] Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

[4] Department of Chemistry, Technical University of Munich, 85748 Garching, Germany

Published online: 12 July 2021

mechanism involves *recombination-like events* (Gemayel et al. 2012) such as unequal crossing-over and gene conversion. Low complexity regions are prone to insertion or deletion mutations. Multiple factors, such as the presence of multiple repetitive units (Legendre et al. 2007), the length of the LCR, and the nucleotide repeat purity of it (Saitou 2018), have been proposed to explain this instability. Conversely, a non-specific use of synonymous codons has been shown to drastically increase the stability of LCRs (Verstrepen et al 2005). Lastly, the nucleotide composition is also crucial, with poly-A or poly-T tracts being more stable than poly G or poly C tracts (Gragg et al. 2002).

Growing evidence stresses the functional role of LCRs in many cellular processes. Shen et al. (2004) showed how the Arginine and Serine-rich binding sites in Exonic Splicing Enhancers contribute to the assembly of pre-spliceosomes, supporting splicing and related activities. Similarly, histidine-rich sites are pivotal for subcellular localization (Salichs et al. 2009). Overall, increasing evidence is accumulating that low complexity regions are preferentially inserted only in specific functional protein classes (Karlin et al. 2002; Albà and Guigo 2004; Faux et al. 2005). As reminded above, *Plasmodia* are a genus of successful parasites, and it is interesting to specifically understand the functional roles of LCRs in their evolution. *P. falciparum* is the etiological agent of the most severe and lethal form of human malaria and its proteins display an abundance of low complexity regions that are rich in asparagine (N) residues (Pizzi and Frontali 2001). Despite the vast literature on *P. falciparum*, the functional implication of its LCRs still remains unclear. In their seminal papers, Pizzi and Frontali (2001) and other authors (Karlin et al. 2002; Ferreira et al. 2003) suggest that plasmodial LCRs are a source of antigenic variations that allow the parasite to evade the host immune response. N-repetitive stretches have been lately proposed, at the genomic level, to influence the local rate of translation, triggering ribosome pausing and ultimately acting as tRNA sponges that assist co-translational folding (Frugier et al. 2010; Filisetti et al. 2013). Alternatively, Forsdyke (2016) discusses the results obtained with Xue (2003), proposing them as intron-like regions stabilizing mRNAs. Muralidharan et al. (2012, 2013) discuss the dispensability of these tracts and how they are targeted by heat shock proteins preventing them to form deleterious amyloid-like fibrils, induced by the thermal variations by which *P. falciparum* is affected, in its life cycle.

In this paper, we investigate whether the LCRs of *Plasmodia* are subjected to a neutral or rather a selective and species-dependent evolution and which factors help the fixation of these regions in plasmodial genomes. We considered the proteomes of 22 *Plasmodium* species, in which we first studied the different correlations between protein length and the abundance of low complexity regions. We observed that *P. falciparum* and *Laverania plasmodia* (Otto et al. 2018), sharing the same GC genomic content and ancestry, show a unique tendency to include LCRs, of specific composition, because of a Darwinian selection. The comparative analysis carried out using the relative synonymous codon usage (RSCU), Shannon entropy, and quantifying the use of single codons in LCRs, further emphasizes that the adaptation of the codon usage bias (CUB) of low complexity regions is parasite-specific, suggesting that LCRs can be thought as genetic fingerprints of plasmodial evolution. In the second part of the work, we have separated the proteins of each species into two operational classes: LCPs, proteins containing low complexity regions, and nLCPs, proteins without low complexity regions (as predicted by the SEG algorithm, see 'Methods'). We then studied the relative contribution of mutational bias and selective pressure in shaping the CUB of LCPs and nLCPs proteins. For this purpose, we devised the *Selective Pressure Index* (*SPI*), based on the Effective Number of Codons (ENC, Wright 1990), to effectively quantify the impact of natural selection on plasmodial proteins. In addition to Parity rule 2 plots (Pr2-plots, Sueoka 1995), the Effective Number of Codons and the SPI indicate that LCPs use a broader set of synonymous codons, and are, as expected, less exposed to selective pressure than nLCPs. Furthermore, we noticed that LCPs are intrinsically longer than nLCPs, pointing to a causative link between protein length and the emergence of LCRs. Overall, our results apply consistently to all *Plasmodia* indicating that low complexity regions, once fixated in a plasmodial genome, are not just relics of either slippage or recombination-like events but tend to acquire a specific role in the biology of *Plasmodium* parasites.

Methods

We have investigated various indicators of codon bias (GC content, relative synonymous codon usage, Shannon entropy, ENC plots, Pr2 plots) to derive information about evolutionary pressures exerted on the genes of plasmodial proteins, with and without LCRs.

Data sources and general statistical methodologies

Plasmodium proteomes were downloaded from the NCBI GenBank (ftp://ftp.ncbi.nih.gov). For *P. vivax* we relied on GCA_000320645.2. We retrieved *P. praefalciparum, P. adleri* and *P. o.curtisi* complete coding sequences (CS) sets from PlasmoDB (https://plasmodb.org). We considered only coding sequences (i) starting with the AUG codon and ending with one of the stop codons (UAG, UAA, UGA); (ii) having a length in base-pairs that is a multiple of three; (iii) without unidentified bases. Proteome numbers are collected in SM (Table SM.1).

Each CS was translated in the corresponding amino acid sequence using the *nt2aa* MATLAB function. We often used a boxplot visualization strategy: considering two boxplots, the MATLAB notch function indicates, when two notches do not overlap, that medians of boxes differ with 95% confidence (Mathworks). Best fit models were calculated with the MATLAB *fit* function. Most of the other analyses were done using in-house MATLAB scripts, available, upon request, from the corresponding author.

We used SEG (Wootton and Federhen 1996) algorithm with W = 15, K_1= 1.5, and K_2= 1.8, to identify LCRs strongly polarized towards a certain species of amino acids, whilst allowing LCRs with a more heterogeneous repertoire of amino acids to be found (Radó-Trilla and Albà 2012). Following SEG predictions, we stratified the proteomes of the *Plasmodium* parasites by dividing them into two operational classes: LCPs, proteins containing low complexity regions, and nLCPs, proteins without low complexity regions. For each *Plasmodium* species, we compared the length distributions of nLCPs and LCPs. The length of each protein was measured using the MATLAB *length* function applied to nucleotide sequences.

For statistical tests, we used the Welch-t-test when the shape of the distributions does not considerably diverge from normality, otherwise, the Mann Whitney *U* test was used. For multi-comparison tests we relied on Welch ANOVA, followed by Bonferroni correction when the distributions did not considerably deviate from normality (*Multiple Comparison Tests 1* (MT1)). Otherwise, the Kruskal–Wallis test, followed by Bonferroni correction was denoted as MT2. For each of these multi-comparison tests, we refer to a MATLAB interactive plot, provided in SM, to visualize and compare each distribution considered in the test, thus allowing the reproduction of Bonferroni corrections.

Subgenera

In this work, we have focused on 22 *Plasmodium* species able to infect a range of different vertebrate organisms. Below, we provide a brief description of these parasites and their phylogenetic relationships. *P. falciparum* has been placed together with the other parasites belonging to the *monophyletic subgenus* termed *Laverania* (Otto et al. 2018). This subgenus comprises *P. gaboni* and *P. reichnowi* (that infect chimpanzees), *P. praefalciparum* and *P. adleri* (that infect gorillas). Noteworthy, *P. falciparum* successfully infects humans (Otto et al. 2018) but, even though it was for a long time considered a human-specific pathogen, it also infects gorillas, raising concerns about possible reciprocal host transfer (Prugnolle et al. 2010). As far as the Asian monkey parasites are concerned, we retrieved data for *P. vivax*, representing a serious threat for human health (Howes et al. 2016), *P. knowlesi,* recently recognized as a human infecting parasite (Rich and Xu, 2011), *P. cynomolgi,* infecting old world monkeys such as baboons and macaques, (taxonomically *Cercopithecidae*), *P. coatneyi* , infecting *M. mulatta*, *P. inui,* infecting *M. assamensis* and *M. fascicularis, P. fragile* and *P. gonderi* (Arisue et al. 2019). We refer to these parasites as S*imian Plasmodia*. To have a blueprint about qualitative and quantitative diversification of LCRs we also considered murine rodents *Plasmodia* (*Vinckeia Subgenus*). We considered *P. vinckei* (Carter and Wallinker 1975), *P. petteri, P. chabaudi, P. berghei* and *P. yoelii* (Garnham 1964). The initial alleged evidence for a descendance of *P. falciparum* from avian *Plasmodia* (Waters et al. 1993a, b) has been lately questioned due to the small number of ingroup taxa and the marker chosen in the early study, namely 18SrRNA (Escalante and Ayala 1994). To keep in line with these former studies, we extended our dataset with the two available specimens of what we refer to as the *Haemamoeba Subgenus* (Corradetti et al. 1963): *P. gallinaceum*, and *P. relictum*, with the latter being one of the most geographically widespread malaria parasites for birds (Valkiunas 2000). Lastly, we considered *P. ovale wallikeri* (*P. ovale*), the etiological agent of tertian malaria, (Collins and Jeffery 2005), *P. ovale curtisi* (*P. o.curtisi*) (Kristan et al. 2019), and *P. malariae,* causing quartan malaria (Collins and Jeffery 2007). We refer here to these parasites as *Human Infectious Plasmodia* (*HIPs*).

GC-content

The GC-content of a gene is the percentage of guanine and cytosine bases with respect to the total length of that gene. Likewise, it is possible to define the GC-content in the first (GC_1), second (GC_2), and third (GC_3) codon positions, as follows:

$$GC_{1,2,3} - \text{content} = \frac{C_{1,2,3} + G_{1,2,3}}{A_{1,2,3} + T_{1,2,3} + C_{1,2,3} + G_{1,2,3}}$$

RSCU

The relative synonymous codon usage is the observed frequency of a codon divided by the expected frequency if all the synonymous codons for the amino acid were used equally (Sharp and Li 1987). The RSCU is computed for each codon of each amino acid, and it is formally defined as follows. For an amino acid i, let n_i denote the number of synonymous codons encoding for the amino acid i. For the $j-th$ codon of amino acid i, let X_{ij} denote the number of occurrences of the codon j. Then the RSCU of codon j of amino acid i ($RSCU_{ij}$) is determined using the following formula:

$$RSCU_{ij} = \frac{X_{ij}}{\sum_{j=1}^{n_i} X_{ij}} / \frac{1}{n_i},$$

where the summation is taken over all synonymous codons of amino acid i. RSCU is a real value between 0 and the number of synonymous codons for that amino acid (i.e., n_i). If the RSCU is close to 1, synonymous codons are used without apparent biases. RSCU values greater or less than 1 indicate that the corresponding codons are over or underused, respectively. We modified the original version of RSCU proposed by Sharp and Li (1987), in line with the division of the 6-fold codon families into 4-fold and 2-fold proposed by Sun and colleagues (2013). The two single codons for methionine (ATG) and tryptophan (TGG) were excluded in the calculation.

Shannon entropy

To compare the codon heterogeneity of LCRs we relied on *Shannon Entropy* (Shannon 1948) of a LCR computed as the sum over the 61 codons that code for amino acids:

$$H_j = -\sum_{i=1}^{61} p_{ij} \log(p_{ij})$$

where $p_{ij} = \frac{n_{ij}}{L_j}$ is the number of occurrences of the i-th codon in the j-th LCR of length L_j. $\mathrm{H_j}$ measures the heterogeneity in the use of codons in the j-th LCR in a sample. The broader the set of codons that is used along a low complexity region the higher is its entropy.

The effective number of codons

We calculated the Effective Number of Codons to estimate the codon usage bias of *Plasmodium* genes encoding for LCPs and nLCPs genes. The values of ENC range from 20, when just one codon is used for each amino acid, to 61, when all the synonymous codons are equally used for each amino acid (Wright 1990). Therefore, the smaller the ENC value, the larger the extent of codon preference in a gene. To calculate ENC values of all the individual genes ($\geq 300bp$) we used the improved implementation by Sun and co-workers (2013). The six-fold codon families (Leu, Ser and Arg) were divided into two-fold and four-fold codon families. We quantify Fα for each coding sequence, defined for each synonymous codon family α as:

$$F_\alpha = \sum_{k=1}^{m_\alpha} \left(\frac{n_{k_\alpha}}{n_\alpha} \right)^2,$$

where m_α is the number of codons in the codon family α, $n_{i\alpha}$, with $i = 1, 2, \ldots, m_\alpha$, is the number of occurrences of the i-th codon in the codon family α and $n_\alpha = \sum_{k=1}^{m_\alpha} n_{k_\alpha}$. Finally, the gene specific ENC is defined as:

$$ENC = N_s + \frac{K_2 \sum_{\alpha=1}^{K_2} N_\alpha}{\sum_{\alpha=1}^{K_2} \left(n_\alpha F_\alpha\right)} + \frac{K_3 \sum_{\alpha=1}^{K_3} N_\alpha}{\sum_{\alpha=1}^{K_3} \left(n_\alpha F_\alpha\right)} + \frac{K_4 \sum_{\alpha=1}^{K_4} N_\alpha}{\sum_{\alpha=1}^{K_4} \left(n_\alpha F_\alpha\right)},$$

where N_s is the number of codon families with a single codon (i.e. Met and Trp), and K_m is the number of families with degeneracy m.

ENC plot

We implemented an ENC plot analysis to assess the relative contributions of mutational bias and natural selection in shaping the codon usage bias of *Plasmodium* genes. The ENC plot is a plot in which the ENC is the ordinate and the GC_3 is the abscissa (Forcelloni and Giansanti 2020). Depending on the action of mutational bias and natural selection, different cases are discernible. If a gene is not subject to selection a clear relationship is expected between ENC and GC_3 described as follows:

$$ENC = a + b \cdot GC_3 + \left\{ \frac{c}{GC_3^2 + d \cdot \left(1 - GC_3^2\right)^2} \right\}$$

where $a = 1$, $b = 2$, $c = 29$ and $d = 1$. Genes for which the codon choice is neutral are expected to lie on or just below Wright's theoretical curve. Conversely, if a gene is subject to selection, it will fall below the theoretical curve. In this case, the vertical distance that separates the gene (represented by a point) and the curve provides an estimation of the relative extent to which natural selection and mutational bias affect the codon usage bias (Forcelloni and Giansanti 2020).

Selective pressure index

The distance from Wright's theoretical curve provides an estimate of the extent to which mutational bias and natural selection affect the codon bias of that gene (Novembre 2002). Depending on the GC_3 value, Wright's theoretical curve establishes the range of variation of ENC values of genes. For this reason, similar distances measured for different GC_3 values are not equivalent but correspond to different extents of selective pressure. To solve this misconception, we propose here the *Selective Pressure Index* defined as the ratio between the observed distance of a gene from Wright's theoretical curve and the maximum distance expected according to the GC_3 of that gene:

$$SPI\left(GC_3\right) = \frac{ENC_w\left(GC_3\right) - ENC_p\left(GC_3\right)}{ENC_w\left(GC_3\right) - e},$$

where GC_3 is the average GC-content in the third codon position of the gene; $ENC_w(GC_3)$ is the value of Wright's theoretical curve for that GC_3 value; $ENC_p(GC_3)$ is the ENC value for the gene in correspondence of that GC_3 value; e = 20 corresponds to the case of extreme bias when just one synonymous codon is used for each amino acid. *SPI* weights the shift of the coding sequence from the null model of no-codon preference (the situation where a gene lies on the curve) with the situation of extreme bias, namely where just one codon is used for each amino acid, for that GC_3. When a gene lies on the theoretical curve SPI is expected to be 0 (no distance from the WTC). Otherwise, when a gene is extremely subject to natural selection SPI is expected to be 1 (situation of extreme bias 20 codons for 20 amino acids).

Pr2 plot

We performed a Parity Rule 2 plot analysis to assess the relative contribution of mutational bias and natural selection on CUB of genes encoding for nLCPs and LCPs through the lens of 4-fold codon families. In these plots, the $GC-bias$ $[G_3/(G_3 + C_3)]|_4$ and the $AT-bias$ $[A_3/(A_3 + T_3)]|_4$ at the third codon position of the four-fold degenerate synonymous codon families are plotted as the abscissa and the ordinate, respectively (Sueoka 1995; Sueoka and Kawanishi 2000). Here, '| 4' denotes the four-codon amino acids, and A3, T3, C3, and G3 are fractions of the nucleotides at the third codon position. If data points are located around the center ($G_3 = C_3$, then the mutational bias is the predominant factor shaping codon usage of genes (Forcelloni and Giansanti 2020). Conversely, deviations from the center indicate the action of natural selection in shaping the codon usage bias of genes. Consistently with the other analyses, we applied the codon family stratification proposed by Sun and colleagues (2013).

Results

On the GC-content, length of the proteins and abundance of low complexity regions

We firstly investigated whether the GC content is correlated, in the genomes of *Plasmodium* parasites, with the abundance of LCRs. The mean GC content of the coding regions, with standard deviations, are listed in the first column of Table 1, whereas the percent incidence of LCRs is reported in the third column of the same table. Even at a first glance, we would exclude a universal correlation between GC content and the inclusion of LCRs. Indeed, consider *Laverania* species. These *Plasmodia* show the higher incidence of LCRs in tandem with a low GC content. Nevertheless, AT-rich *Plasmodia* (e.g., *Haemamoeba* and *Vinckeia* parasites) display an incidence of LCRs similar to that of *Simians plasmodia,* which are richer in GC content. The middle column of Table 1 collects the coefficients of the correlation between the extent of LCRs and the length of the proteins, a theme that will be further discussed below. Interestingly, we see that *Laverania* species display stronger correlations than the other subgenera. These rough observations seem to indicate that it is not straightforward to find rules of thumb for the evolutionary, adaptive role of the GC content in *Plasmodium* species, as pointed out by a recent study, that stresses the peculiarity of the individual adaptive trajectory of each species, a matter that is still largely to be investigated (Castillo et al. 2019). After the first exploration, for the sake of completeness, we applied statistical tests to the data summarized in Table 1. Specifically, we applied the MT1 procedure to these distributions (see above: data sources and general methodologies) to test the null hypothesis that all *Plasmodium* genomes have the same GC content. Out of several multiple comparisons, we could assess that, consistently with other works (Videvall 2018), *Haemamoeba plasmodia* have the lowest GC content (rejecting the null hypothesis with $p < 0.01$) whereas most *Laverania* and *Vinckeia plasmodia* have similar GC contents (the null hypothesis is not rejected, with $p > 0.05$). Noteworthy, *Simian plasmodia* have the most GC-rich genes ($p < 0.01$) whilst the *HIPs'* subgroup is placed halfway between *Vinckeia* and *Simian plasmodia,* differing from both ($p < 0.01$). As mentioned above, we then considered the question of whether there is a (possibly universal) correlation between protein length and the abundance of LCRs. The second column of Table 1 collects, for each parasite, the Pearson correlation coefficients of the 22 LCR protein abundance (y-axis) vs protein length (x-axis) distributions. All the correlations were significant (null hypothesis of a zero-correlation rejected with $p < 0.01$). Since the distributions of the r values are not normal, we applied the multiple MT2 protocol to test the null hypothesis that in all the species there is the same correlation between protein length and LCRs. We find significant differences between *Laverania, Vinckeia,* and *Simian plasmodia* ($p < 0.01$), whereas the differences between *Laverania, Haemamoeba,* and HIPs were not significant ($p > 0.05$). We applied the same protocol to the third column of Table 1 where we collected the proportions of LCRs in the proteomes of each parasite. We observed the same statistical differences we obtained for the correlation coefficients. Based on these initial observations, the genomic GC content does not appear to be the driving force for the stabilization and abundance of LCRs in *Plasmodia*, in accordance with more articulated previous results (Castillo et al. 2019).

Table 1 The first column, *Plasmodium* species

Organism	GC content	Pearson-correlation coefficient	Percent LCR content
Laverania subgenus			
**P. falciparum*	0.25 ± 0.041	$r = 0.70$	130020/4184097 (0.031)
P. praefalciparum	0.26 ± 0.045	$r = 0.71$	125491/4452734 (0.030)
P. reichnowi	0.25 ± 0.047	$r = 0.71$	133878/4269716 (0.032)
P. adleri	0.24 ± 0.040	$r = 0.67$	105572/4124122 (0.027)
P. gaboni	0.24 ± 0.042	$r = 0.70$	102893/3796446 (0.027)
Simian plasmodia			
**P. vivax*	0.42 ± 0.09	$r = 0.35$	28753/4024747 (0.0071)
**P. knowlesi*	0.40 ± 0.04	$r = 0.26$	19693/3713699 (0.0053)
P. coatneyi	0.42 ± 0.04	$r = 0.32$	15273/3863819 (0.004)
P. cynomolgi	0.40 ± 0.06	$r = 0.42$	26101/3290064 (0.008)
P. inui	0.43 ± 0.047	$r = 0.26$	16109/3768395 (0.0043)
P. fragile	0.42 ± 0.054	$r = 0.22$	19378/3927950 (0.0049)
P. gonderi	0.30 ± 0.04	$r = 0.39$	31857/4055981 (0.008)
Vinckeia subgenus			
P. yoelii	0.26 ± 0.07	$r = 0.33$	29765/3400464 (0.009)
P. berghei	0.25 ± 0.04	$r = 0.38$	28753/3417544 (0.0084)
P. petteri	0.26 ± 0.04	$r = 0.21$	11737/3415448 (0.0034)
P. vinckei	0.26 ± 0.04	$r = 0.16$	11324/3340277 (0.0034)
P. chabaudi	0.27 ± 0.04	$r = 0.26$	11265/3485217 (0.0032)
Haemamoeba subgenus			
P. gallinaceum	0.22± 0.04	$r = 0.46$	49707/3737574 (0.013)
P. relictum	0.23± 0.04	$r = 0.37$	39707/3659322 (0.0109)
Human infectious plasmodia (HIPs)			
**P. malariae*	0.30 ± 0.04	$r = 0.45$	31876/4055981 (0.008)
**P. ovale*	0.34 ± 0.06	$r = 0.43$	43101/4501484 (0.0071)
P. o.curtisi	0.32 ± 0.045	$r = 0.30$	32005/4423193 (0.0097)

The second column, mean GC content ± std of plasmodial proteins. The third column, Pearson correlation coefficients between LCR abundance and protein lengths. The fourth column, the proportion of LCRs, calculated as the ratio of the number of residues that belong to LCRs to the total number of residues in the proteome. Note, specifically, the higher incidence of LCRs in *Laverania*. Note, also, the remarkable shift of GC content in *Simian plasmodia*. Interestingly, in Laverania there is a marked correlation between the length of the proteins and the propensity to accommodate LCRS. Asterisks indicate species that are known to cause malaria in humans

Codon bias and LCRs

We studied the codon usage bias of LCRs using the relative synonymous codon usage (Sharp and Li 1987). In Fig.1, we show a heatmap of the RSCU values of LCRs. This data is computed on the codon content of the LCRs of each *Plasmodium* species. Data is standardized along each column. If a codon has a value above the column mean, it is represented in red. Otherwise, it is shown in green. Details of the clustering algorithm and distance are provided in the figure caption. The row dendrogram indicates that *Plasmodium* parasites can be separated into two main groups. *Laverania, Haemamoeba,* and *Vinckeia plasmodia* emerge at the bottom of the heatmap, characterized, as expected, from the data in Table 1, by enrichment in AT-ending codons. *HIPs* and *Simian Plasmodia* are placed at the top and are characterized by enrichment in GC-ending codons. The row-dendrogram, on the left side of the heatmap, returns consistent phylogenetic relationships: *P. cynomolgi* and *P. vivax* are closer than the other *Simian plasmodia* species (Tachibana et al. 2012); *P. gonderi* is placed in the same lineage of *Simian plasmodia* (Arisue et al. 2019), pointing to common ancestry and similar selective pressure in the choice of synonymous codons of their LCRs. Remarkably, the part of the dendrogram relating to the AT-rich *Plasmodia* also returns satisfactory results, as it is respected the phylogenetic relationships of the *Laverania*, *Haemamoeba,* and *Vinckeia plasmodia* highlighted in other works (Valkiunas et al. 2018; Larson 2019). A consistent phylogeny is also reported for *P. ovale* and *P. o.curtisi* (Kristan et al. 2019). To complement these observations, we

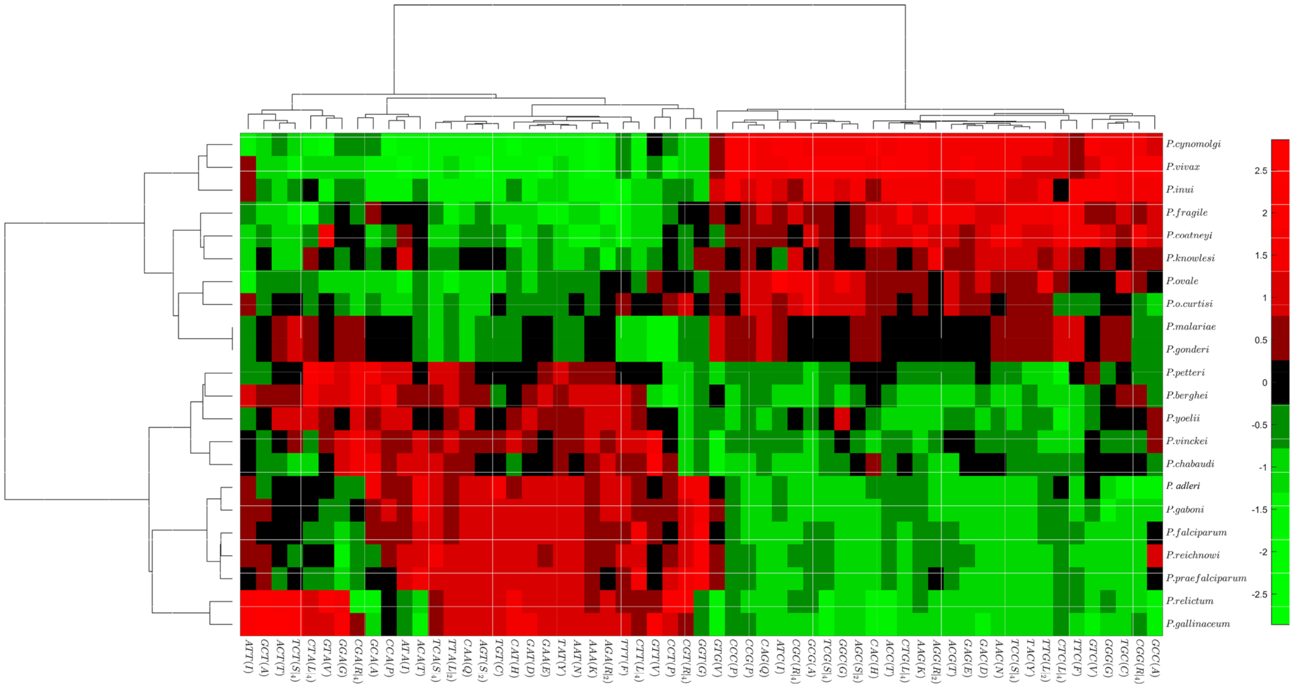

Fig. 1 Clustering of the relative synonymous codon usage (RSCU) vectors associated with each species. In each row, we show a different Plasmodium species. Each column corresponds to a codon. Data is standardized along the columns. Ward's Linkage was used. We used MATLAB clustergram function. Row and column pairwise distance was calculated with Euclidean Distance. In red, we show codons with a value in the species reported in the row which is greater than column average. In green, we show codons with a value in the species reported in the row which is lower than column average

compared the RSCU values of codons ending with a purine (A or G) and codons ending with a pyrimidine (T or C) (Mann Whitney test of the null hypothesis that purine ending and pyrimidine ending codons have the same RSCUs). The RSCU is a normalized index that ranges between 0 and the degeneracy of the codon family under study. To sensibly compare different codon families, we normalized the RSCU values with respect to the degeneracy of the codon family they belong to. In line with what is observed in the heatmap, *Laverania, Vinckeia,* and *Haemamoeba plasmodia* show a preference for A over G and T over C, in the third codon position ($p < 0.01$). The heatmap has highlighted how *HIPs* have RSCU values, related to codons with an Adenine or a Thymine in the wobble position, lower than *Vinckeia*, *Haemamoeba,* and *Laverania plasmodia*. Nevertheless, the highest RSCU values for *HIPs* are those for codons with an A or T in the wobble position ($p < 0.01$). We find the same pressures in *P. gonderi* ($p < 0.01$). Interestingly, we do not find significant differences in many of the *Simian plasmodia* (*P. vivax, P. cynomolgi, P. fragile,* and *P. inui*), whose wobble position appears to be caught in a '*tug-of-war*' between purines and pyrimidines, without significant differences ($p > 0.05$). Evolutionary pressures are slightly different in *P. coatneyi* and *P. knowlesi*. The first prefers G over A and T over C ($p < 0.05$). The second displays wobble positions consistent with its GC content by preferring G to A and C to T ($p < 0.05$). The general RSCU table is provided in SM (see *RSCU.xlsx*). Overall, despite the similarities that unite different parasitic sub-groups, *Plasmodium* species show diversified tendencies to use synonymous codons as can be furtherly observed from the averaged RSCU codon values of the heatmap. Furthermore, the possibility to reconstruct consistent phylogenetic relationships from LCRs points to a link between LCRs and protein evolution.

Different codon repertoires in the LCRs of different *Plasmodium subgenera*

We investigated how the codon composition of LCRs varies in *Plasmodium* species, using Shannon entropy. We calculated the entropy of each low complexity region, which is a measure of how broad its codon repertoire is. In Fig.2a we reported the distributions of the five plasmodial groups (defined in the methods). Looking at the regression lines, for the same length, the LCRs of *Laverania plasmodia* are less complex (relatively lower entropy) than those of the other parasitic groups. This means that LCRs of *Laverania plasmodia* use a narrower set of codons than the other *Plasmodia*. Next, we considered the distributions of the LCRs entropies in each group of *Plasmodium* species, as shown

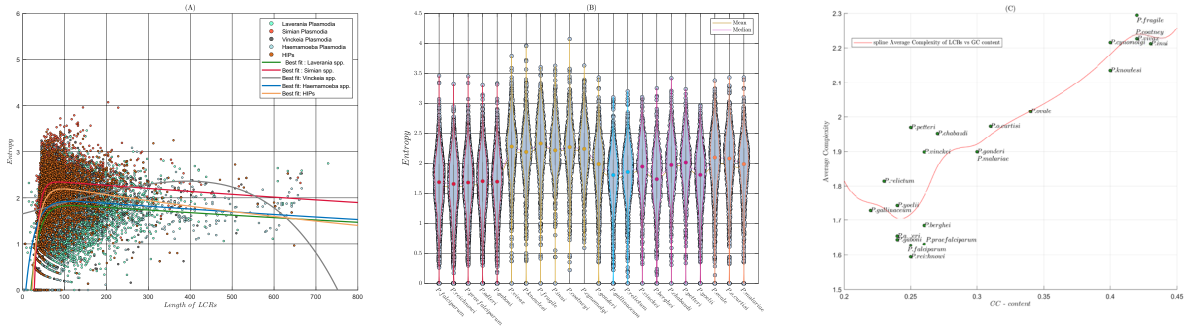

Fig. 2 **a** Regression of Entropy vs LCRs length. Length of LCRs is measured in nucleotides. Model parameters are provided with 95% coefficient bounds Laverania(x) = $a \times e^{b\cdot x} + c \times e^{d\cdot x}$: a = 1.918 (1.885, 1.951), b = − 0.0003328 (− 0.0004543, − 0.0002113), c = − 5.435 (− 5.793, − 5.078), d = − 0.05285 (− 0.05514, − 0.05056);Simian(x) = $a \times e^{b\cdot x} + c \times e^{d\cdot x}$: a = 2.41 (2.368, 2.452), b = − 0.0002985 (− 0.0004451, − 0.0001519), c = − 39.41 (− 47.79, − 31.02), d = − 0.1012 (− 0.1077, − 0.09478); Vinckeia(x) = $a \times e^{b\cdot x} + c \times e^{d\cdot x}$: a = − 1.006e+04 (− 4.838e+12, 4.838e+12), b = 0.002788 (− 52.56, 52.57),c = 1.006e+04 (− 4.838e+12, 4.838e+12),d = 0.002788 (− 52.55, 52.56); Haemamoeba(x) = $a \times e^{b\cdot x} + c \times e^{d\cdot x}$: a = − 2.832 (− 3.196, v2.469),b = − 0.04565 (− 0.05072, − 0.04057),c = 2.018 (1.956, 2.08), d = − 0.0003467 (− 0.0004884, − 0.000205); HIPs(x) = $a \times e^{b\cdot x} + c \times e^{d\cdot x}$: a = − 9.49 (− 10.81, − 8.165), **b =** − 0.06278 (− 0.0674, − 0.05817), **c =** 2.357 (2.291, 2.423), d = − 0.0006497 (− 0.0008514, − 0.0004479). **b** Illustration of the average complexity of each Low Complexity Region. The mean of each distribution is represented by a data point of the same colour as the outline of the violin plot. The trend of the means and medians is represented by the yellow and red lines respectively. Violin Plot Function has been taken from GitHub-Matlab. **c** Trend of the sample averages of the distributions of **b** with respect to the average Guanine and Cytosine content of each parasite (Table 1)

in Fig.2b. Visual observation of the violin plots shows that there is a species-specific signal. The diverse subgenera differently modulate the codon content of their LCRs, which denotes, as entropy increases, a greater variety of codons occurring along a LCR. Given the observed symmetry of the violin plots, we applied the MT1 multiple test protocol for differences between the distributions of different groups. The Bonferroni corrected statistical significances confirm what has been observed through the regressions, indicating a lower codon heterogeneity of LCRs in *Laverania subgenus,* if compared to the other groups of *Plasmodia* (p ≪ 0.01). The comparison between *P. berghei* and the *Laverania* group is, however, not significant (p > 0.05). Fig.2c shows the correlation between the average codon complexity of LCRs and the average GC content of each parasite (listed in Table 1). The spline indicates that as the GC content increases, the codon heterogeneity of the LCRs increases as well. However, the fluctuations further confirm that there are other factors besides GC content that influence the codon repertoire of LCRs. Noteworthy, further information is reported in SM (Fig. SM.12–16). Refer to these graphs for a better understanding of the codon composition of LCRs and of the results that are shown in this paragraph.

Quantification of codon usage bias

The RSCU is an index that is normalized internally, with respect to each family of synonymous codons. It should be weighed considering the relative frequencies with which the amino acids occur in a protein. In Fig. 3 each bar (P_{Cdn}) represents the ratio between the total number of a codon in the LCRs over the total number of codons that compose LCRs, in each parasite. Each bar plot contains information relating to a subgroup of *Plasmodia*, as defined in the '*Subgenera*' paragraph. For each codon, there are as many bars as there are *Plasmodia* that make up the subgroup, respectively. From a first visual observation of the bar charts, and as a rule, we find that the *Plasmodium* species of the same group have overlapping distributions, which intuitively reinforces the hypotheses we have advanced through the RSCU analysis. Deepening our investigation, codons in LCRs translate, in essence, for the same amino acids which are N, E, D, S and, K. This is reported in all the analyzed species. Worth noting, N, E, D, S and, K have similar chemical-physical properties, being, moreover, their reciprocal most common substitutions (NCBI—Amino Acid Explorer). Therefore, the composition of LCRs suggests a conservative amino acid selection (Strachan et al. 2019). Regardless of the similar preferred amino acid patterns, the complexity of the codon repertoire increases dramatically in *HIPs* and *Simian plasmodia*. Even the AT-rich species exhibit a certain variability in codon usage, although drastically less intense than the more GC-rich *Plasmodia*. *Laverania subgenus* shows the narrowest set of utilized codons.

LCPs are longer than nLCPs

We compared the length distribution of LCPs and nLCPs in each *Plasmodium* (Fig. 4a). LCPs are significantly longer than nLCPs in each parasite (Welch-t-test, p ≪ 0.01). In

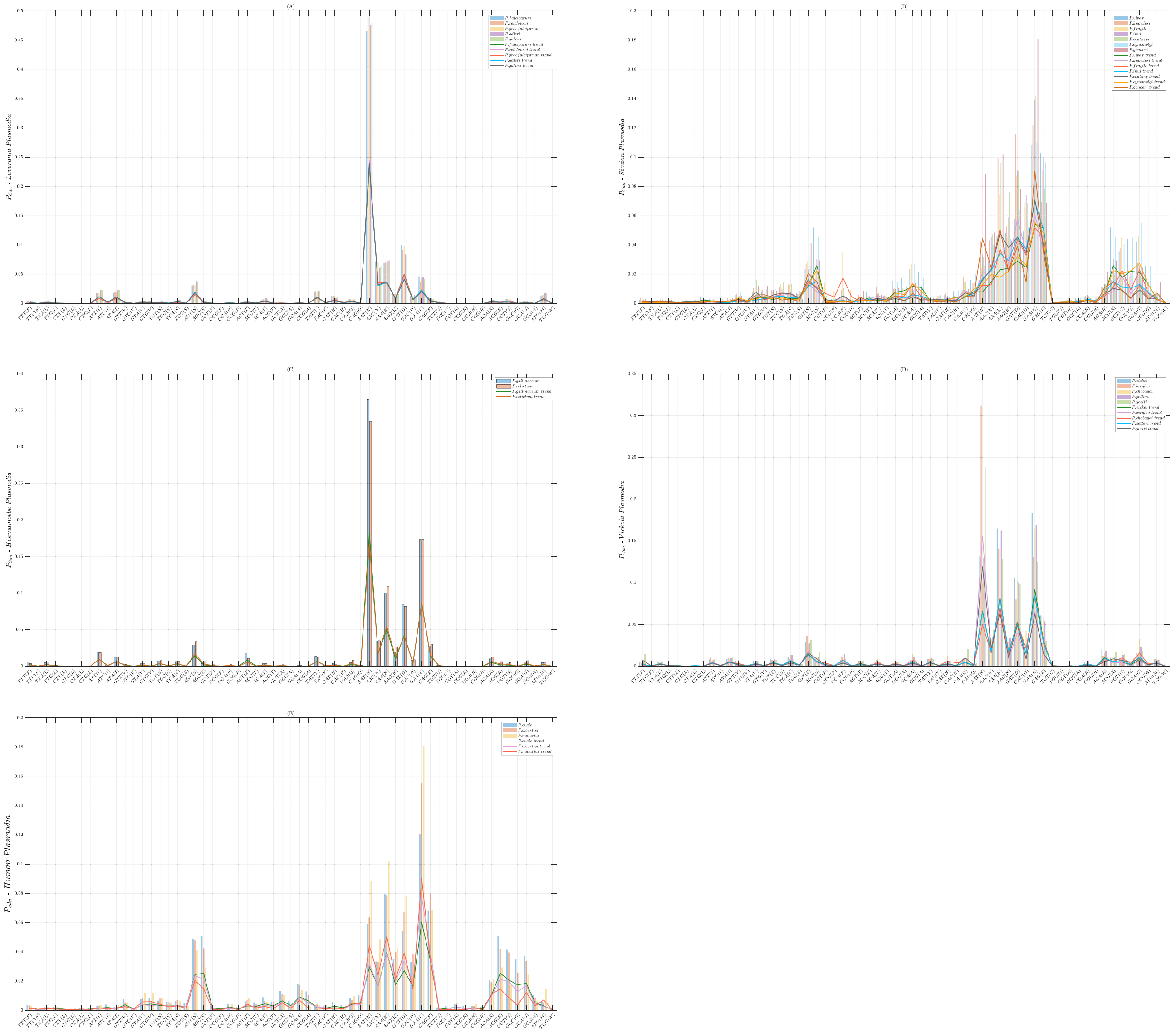

Fig. 3 Percentage representation of the CUB in the various parasitic subgroups. The amount of each codon is normalized over the total amount of codons present in the LCRs of each parasite. **a** *Laverania subgenus* **b** *Simian plasmodia* **c** *Haemamoeba subgenus* **d** *Subgenus vinckeia* **e** HIPs. Each peak represents half of the bar it refers to

general, a single low complexity region does not exceed 250 amino acids of length (see Supplementary Materials Fig SM.11a). However, more than a single low complexity region is present on average in plasmodial proteins (see Supplementary Materials Fig SM.11b). Indeed, LCRs cover 10–20 % of proteins' length, on average (Fig. 4b). Motivated by these observations, we investigated whether the larger size of LCPs might be due to the presence of these stretches. We, therefore, deprived LCPs of their LCRs and repeated the experiment by comparing the length distribution of LCPs and nLCPs, in each parasite (Fig. 4c). Interestingly, LCPs are intrinsically longer than nLCPs (Welch t-test, $p \ll 0.01$). We applied MT1 protocol to the distributions of LCPs, to understand if there could be differences explaining the overabundance of LCRs of *Laverania plasmodia* and if their overabundance was due solely to the length of their proteins. The most relevant information that comes out of MT1 is that *P. gonderi* and *P. malariae* emerge to have the longest LCPs and *P. berghei* appears instead to have the shortest, on average ($p < 0.01$). Therefore, no differences emerge such as to justify the overabundance of LCRs of *Laverania subgenus*. We conclude that the length of the proteins provides a selective advantage for the emergence of low-complexity regions. However, given the disparity between the abundance of LCRs between *Laverania plasmodia* and the other species, we conclude that protein length is a necessary feature for the emergence of these regions but, it is not sufficient to explain the overabundance in *P. falciparum* and its closest relatives,

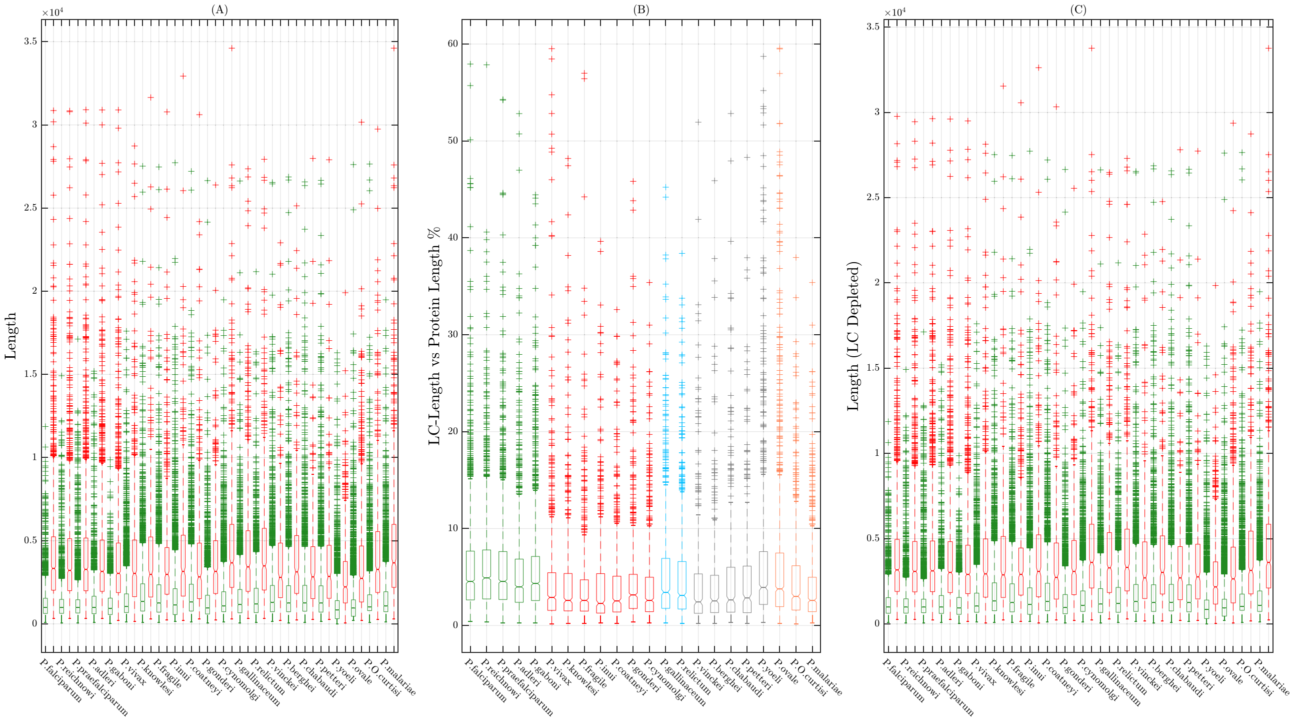

Fig. 4 **a** Protein length comparison of LCPs and nLCPs in each parasite. **b** Percentage ratio between low complexity content and protein length in the proteomes of *Plasmodium* species. LCR length vs Protein length ration. **c** Pairwise comparison of LCPs, deprived of LCRs, and nLCPs. The LCPs are represented in red. The nLCPs are represented in green

which corroborates the overall picture indicating LCRs as species-specific.

ENC plots indicate that the codon usage bias of LCPs is primarily shaped by mutational bias

We performed an ENC-plot analysis to estimate the relative contribution of mutational bias and natural selection in LCPs and nLCPs. *Laverania, vinckeia* and, *Haemamoeba subgenera* (Fig. 5a, c, d) show similar distributions, placed on the left side of the ENC plane. Noteworthy, they display a low GC content in the wobble codon position. Interestingly, the nLCPs of *P. yoelii* mainly group on the left side of the ENC plot. Nonetheless, they follow the general shape of Wright's Theoretical Curve, reflecting the GC_3 variation. *HIPs'* genes (Fig. 5b) have a GC_3 content halfway between the genes of the former subgenera and of the *Simian plasmodia,* whose ENC plots are in line with other works (Yadav and Swati 2012; Gajbhiye et al. 2017). *Plasmodium vivax* and *P. cynomolgi* have a portion of their nLCPs positioned on the left side of the ENC plane (Fig. 6). This corroborates their close phylogenetic relationship (Wellcome—Sanger Institute). Visually, the LCPs of each parasite appear in close proximity to Wright's Theoretical Curve, contrary to what it appears to be for nLCPs. As a rule, all AT-rich *Plasmodia* have linearly shaped ENC vs GC_3 distributions. Therefore, we collected their Pearson correlation coefficients in Table 2, together with regression slopes of nLCPs and LCPs. All the correlations are significant (hypothesis of zero correlation rejected with $p < 0.01$). Then, we compared the distributions of the regression slopes of nLCPs and LCPs (Mann Whitney U test). LCPs regressions are characterized by steeper slopes ($p < 0.01$). Mann Whitney test indicates, consequently, a stronger GC_3 pressure on LCPs. In Fig. 6 we show separately the ENC plots obtained for *Simian plasmodia*. Together, we provided best fit curves obtained through a non-linear fit of Wright's shape. In Table 3 we collected the r^2 of regression curves. Specifically, we used a non-linear fit to show averaged trends that better distinguish differential properties between LCPs and nLCPs. Precisely, if the best fit curve lies on or just below the Wright's curve, then the CUB of the protein class will be expected to reflect the effect of mutational bias. Conversely, if the best fit curve lies below Wright's curve, then the CUB of the protein class will be expected to be more under the control of selection (negative/positive), as generally defined in methods. As a rule, the best fit curve of LCPs is closer to Wright's theoretical curve, meaning that mutational bias is stronger on LCPs than on nLCPs. Models and parameters, calculated with 95% confidence intervals (CI), are provided in the caption of Fig. 6. To reinforce the entire set of analyses done so far, we compared the ENC scores of LCPs and nLCPs in each *Plasmodium*

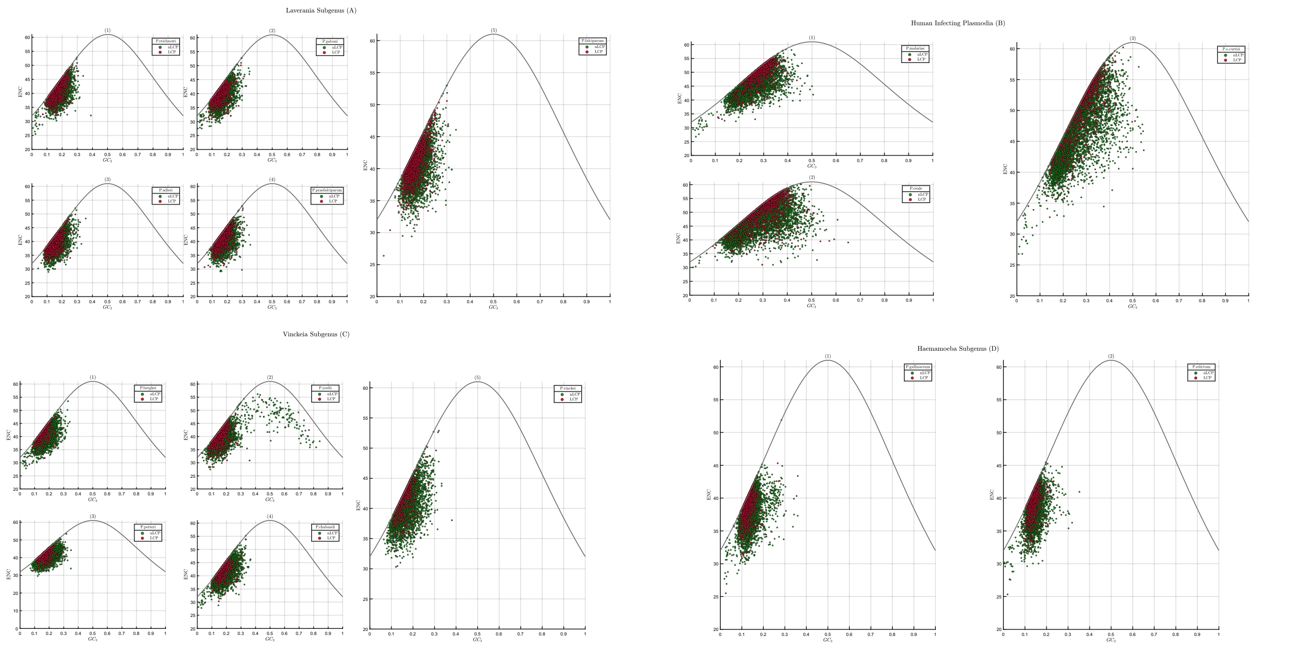

Fig. 5 Illustration of ENC plot for *Laverania subgenus*: (1) *P. reichnowi* (2) *P. gaboni* (3) *P. adleri* (4) *P. praefalciparum* (5) *P. falciparum*; *Human infectious plasmodia*: (1) *P. malariae* (2) *P. ovale* (3) *P. o.curtisi*; *Vinckeia subgenus*: (1) *P. berghei* (2) *P. yoelii* (3) *P. petteri* (4) *P. chabaudi* (5) *P. vinckei*; *Haemamoeba subgenus*: (1) *P. gallinaceum* (2) *P. relictum*. Globally ENC distributions are placed in the left region of the ENC plane. LCPs are represented in red. nLCPs are represented in green

(Fig. 7). Except for *P. chabaudi*, *P. yoelii* and, *P. berghei*, even though the medians of their box plots differ with 95% confidence, all the comparisons are significant with LCPs showing a broader set of synonymous codons with respect to nLCPs (Welch-t-test, p $\ll$ 0.01). Overall, our findings show that the codon usage of LCPs is more relaxed and more affected by mutational bias than the codon usage of nLCPs.

SPI better distinguishes the relative contribution of natural selection and mutational bias

As explained in the '*Methods*' paragraph, the distance from Wright's theoretical curve does not provide an absolute measure of natural selection. This is clear if we consider the same ENC score but different GC_3 values. To overcome this limitation, we introduced the *Selective Pressure Index*. Specifically, we measured the SPI for each gene classified in LCPs and nLCPs operational subclasses. In Fig.8, we show the SPI distributions for LCP- and nLCP- encoding genes. In line with the ENC analysis, LCPs are characterized by lower SPI values than nLCPs. This means that LCPs are closer to Wright's theoretical curve and, therefore, subject to a more relaxed selective pressure than nLCPs. All the comparisons are statistically significant (Welch-t-test, p$\ll$0.01). We evaluated the hypothesis that lower selective pressure, with respect to the other species, could favour a greater pervasiveness of LCRs in the *Laverania subgenus*. We applied the MT1 procedure on the SPI distributions of LCPs. In general, AT-rich *Plasmodia* are comparable (p > 0.05). Therefore, the Bonferroni corrected statistical significances reject the hypothesis that a more pronounced mutational bias favours the LCR abundance of *Laverania plasmodia*. Duret and Mouchiroud (1999) noted a negative correlation between protein length and selective pressure in the proteins of *C. elegans, D. melanogaster,* and *A. thaliana*. We, therefore, decided to retrace what they did using the SPI (Fig. 9). Regression models and parameters, calculated with 95% CI, are provided in the caption. Once more, we deprived LCPs of their LCRs. The graph highlights how the contribution of the mutational bias tends to increase with the length of the protein for both nLCPs and LCPs, which determines a larger set of codons within proteins given the negative linear correlation between ENC and SPI (Table SM.3). Alternatively, this means that selective pressure decreases as protein length increases. The same is reported for the other parasitic groups, whose charts and models are provided in SM (Fig. SM.1–4). Recalling the correlations between LCRs and protein length in Table 1, these trends contain another hidden information: the lower selective pressure on a protein, the more LCRs are inserted.

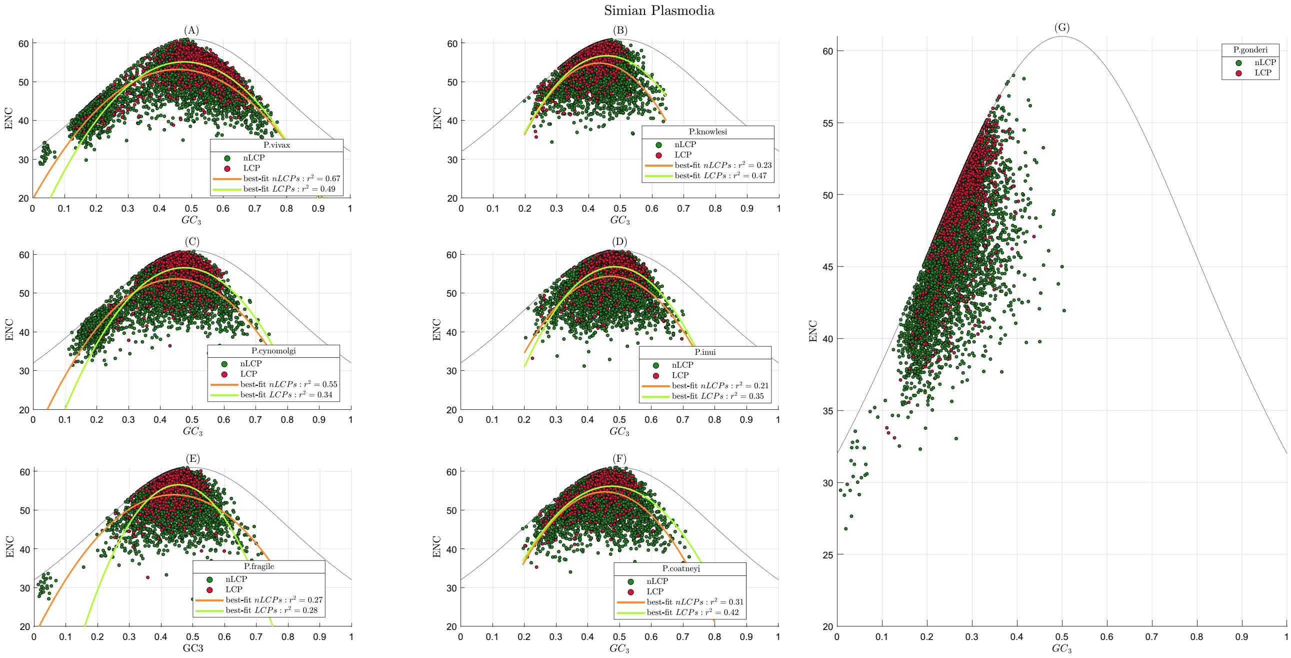

Fig. 6 Illustration of the ENC Plots of *Simian plasmodia*. We used a f(x) = $p_1 \cdot x^2 + p_2 \cdot x + p_3$ model. Best Fit's parameters (BFPs) are provided with the 95% confidence bounds. In red LCPs, in green nLCPs. In orange the best fit for nLCPs. In light green the best fit for LCPs. **a** *P. vivax*: nLCPs' BFPs: p_1 = − 160.3 (− 164.3, − 156.3), p_2 = 146.5 (143.3, 149.7),p_3 = 19.7 (19.13, 20.27);LCPs' BFPs: p_1 = − 197.6 (− 211.5, − 183.7), p_2 = 188.5 (175.5, 201.4),p_3 = 10.18 (7.227, 13.14). **b** *P. knowlesi*: nLCPs' BFPs: p_1 = − 334.7 (− 352.9, − 316.5),p_2= 289.9 (274.4, 305.4),p_3 = − 7.969 (− 11.24, − 4.701); LCPs' BFPs: p_1 = − 298.4 (− 351, − 245.8),p_2 = 272.7 (232.2, 313.2), p_3 = − 5.609 (− 13.32, 2.101) **c** *P. cynomolgi*: nLCPs' BFPs: p_1 = − 207 (− 214.2, − 199.8), p_2 = 185.1 (179.6, 190.7), p_3 = 12.28 (11.23, 13.32); LCPs' BFPs: p1 = − 259 (− 291.5, − 226.5),p2 = 245.9 (217.5, 274.3),p3 = − 1.82 (− 8.019, 4.379) **d** *P. inui*: nLCPs' BFPs: p_1 = − 268.7 (− 284.1, − 253.3), p_2 = 253.2 (239.1, 267.3), p3 = − 5.331 (− 8.551, − 2.111);LCPs' BFPs: p_1 = − 320.5 (− 363.3, − 277.8), p2 = 310.3 (271.1, 349.6),p_3 = − 18.31 (− 27.29, − 9.327) **e** *P. fragile*: nLCPs BFPs: p_1 = − 188.2 (− 197.4, − 179),p_2 = 165.9 (158.1, 173.6),p_3 = 17.42 (15.75, 19.09);LCPs' BFPs: p_1 = − 416.8 (− 476.9, − 356.6),p_2 = 379.3 (327.3, 431.2), p3 = − 29.75 (− 40.93, − 18.57); **f** *P. coatneyi*: nLCPs' BFPs: p_1 = − 280.9 (− 293.3, − 268.6),p_2 = 254.8 (244, 265.6),p_3 = − 3.065 (− 5.406, − 0.724);LCPs' BFPs: p_1 = − 246.7 (− 289.4, − 203.9), p_2 = 234 (198.5, 269.5),p_3 = 0.6524 (− 6.597, 7.902). **g** *P. gonderi*

Table 2 The first column, the names of each *Plasmodium* species

Organism	nLCP	LCP	nLCPs-slope	LCPs-slope
P. falciparum	r = 0.56	r = 0.65	43.9	55.4
P. praefalciparum	r = 0.57	r = 0.62	46.8	51.5
P. reichnowi	r = 0.62	r = 0.62	49.2	52.9
P. adleri	r = 0.61	r = 0.60	47.4	48.1
P. gaboni	r = 0.59	r = 0.53	44.2	41.7
P. yoeli	r = 0.54	r = 0.52	23.4	44.5
P. berghei	r = 0.57	r = 0.62	43.3	51.5
P. petteri	r = 0.54	r = 0.47	41.4	41.5
P. vinckei	r = 0.57	r = 0.54	41.9	43.8
P. chabaudi	r = 0.60	r = 0.51	43.9	39.3
P. gallinaceum	r = 0.42	r = 0.43	30.8	37.8
P. relictum	r = 0.41	r = 0.41	37.7	44.5
P. malariae	r = 0.68	r = 0.72	50.4	63.5

The second and third column, the Pearson correlation coefficients of the ENC distributions of the AT-rich *Plasmodia* of nLCPs and LCPs, respectively. The fourth and fifth column, the slope of the regressions for the nLCPs and LCPs ENC distribution, respectively

Table 3 ENC vs GC_3 r^2 values of GC-rich *Simian plasmodia*

Organism	nLCP	LCP
P. vivax	*r^2= 0.67*	*r^2= 0.50*
P. knowlesi	*r^2= 0.23*	*r^2= 0.47*
P. inui	*r^2= 0.20*	*r^2= 0.35*
P. cynomolgi	*r^2= 0.56*	*r^2= 0.34*
P. coatneyi	*r^2= 0.32*	*r^2= 0.42*

Given the different shape of the distributions present in the ENC plots of *Simian plasmodia*, we used quadratic models to describe their fits. As mentioned in the text of the results, the models and their parameters are shown in the caption of Fig. 6

Pr2 violation is smaller on LCPs than nLCPs

We performed a Pr2-plot analysis for each *Plasmodium* species (Fig.10). LCPs and nLCPs are plotted, respectively, in red and green. As mentioned in methods, we considered four-fold codon families, applying the stratification proposed by Sun et al. (2013). *Laverania*, *Vinckeia,* and *Haemamoeba*

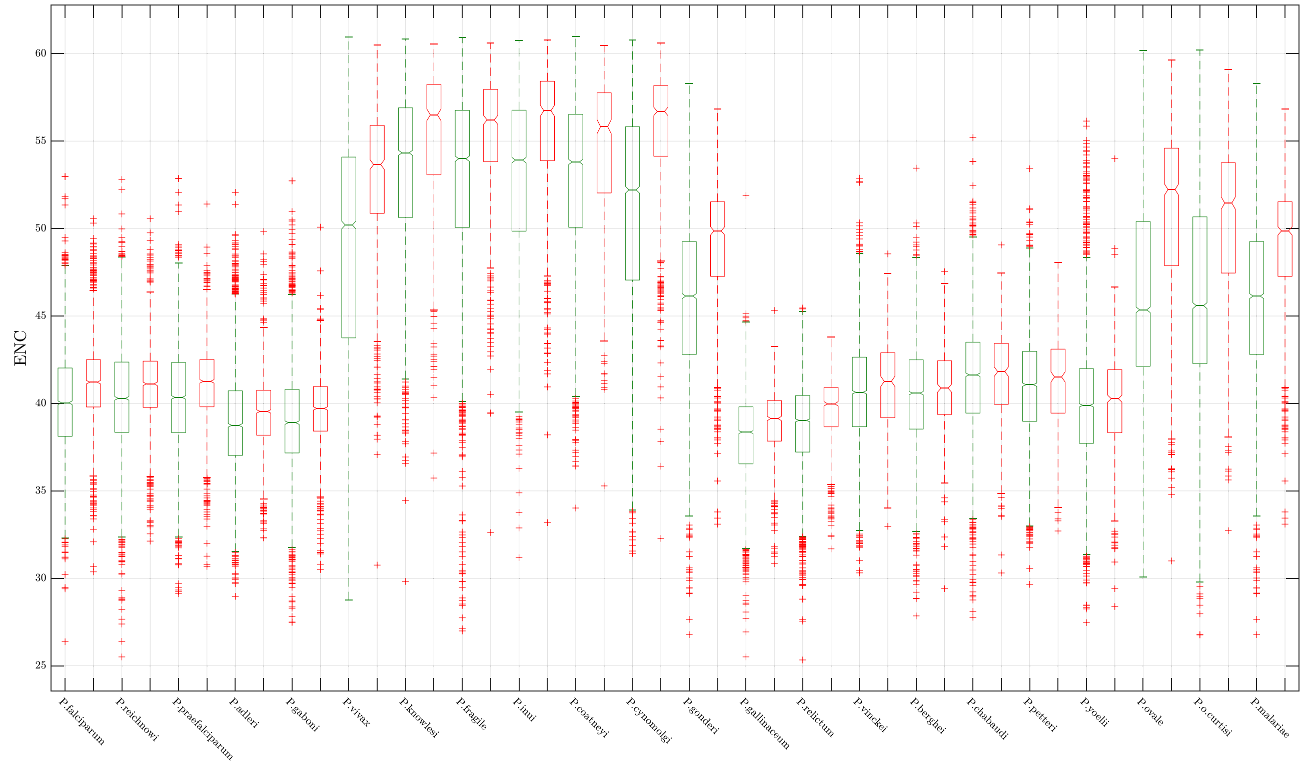

Fig. 7 Pairwise (Welch-t-test) comparison of the ENC distributions of LCPs and nLCPs in each parasite. Consistent with what was done in the previous graphs, the LCPs are represented in red. Similarly, nLCPs are represented in green. The medians of each boxplot pair differ with 95% statistical significance. (Mathworks)

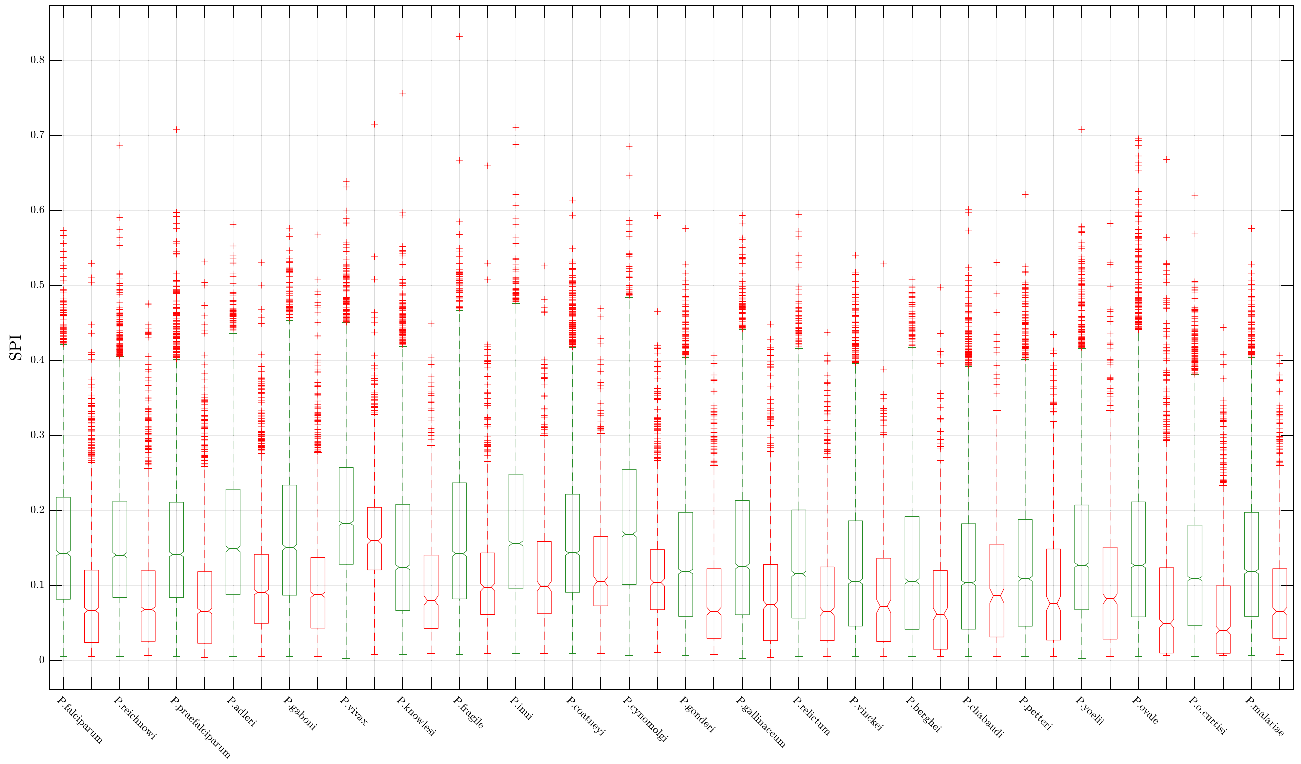

Fig. 8 Pairwise (Welch-t-test) comparison of the SPI distributions of LCPs and nLCPs in each parasitet LCPs are represented in red. In green nLCPs. The medians of each boxplot pair differ with 95% statistical significance. (see Mathworks for the explanation of notch function)

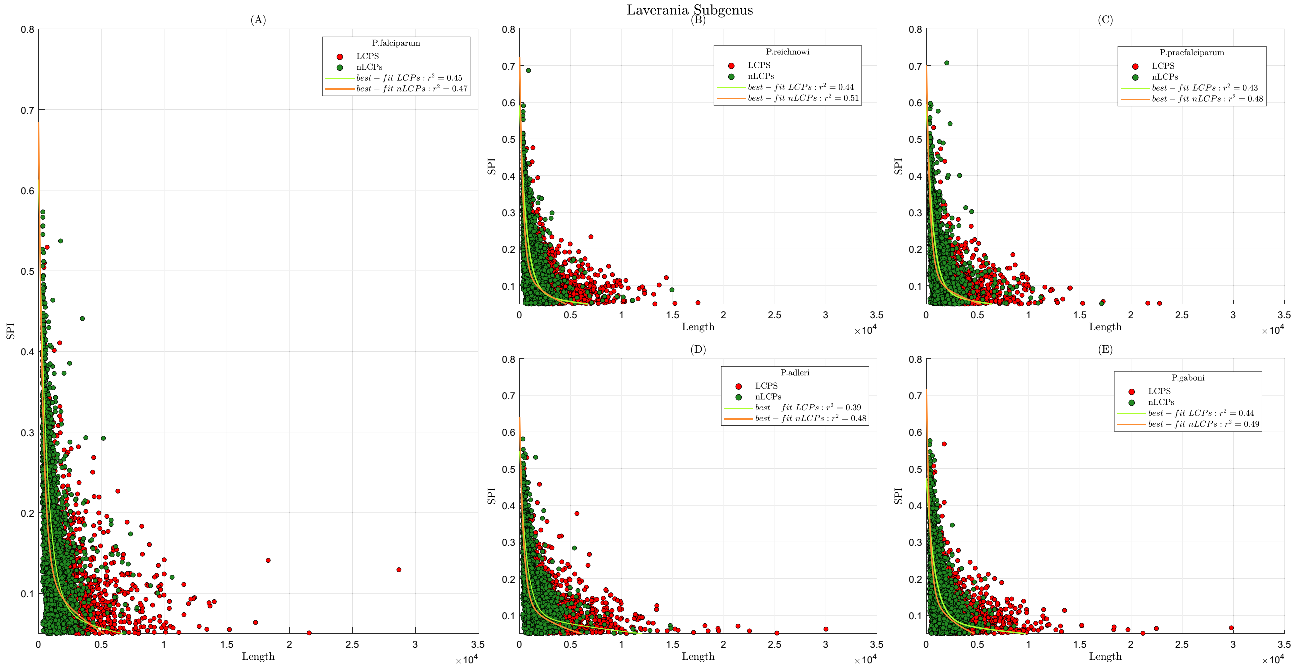

Fig. 9 Illustration of the SPI vs length analysis performed with Laverania parasites. In red LCPs. In green nLCPs. Model: $a \cdot e^{b \cdot x} + c \cdot e^{d \cdot x}$. **a** *P. falciparum*: nLCPs' best fit parameters (BFPs) : a = 0.55 (0.4898, 0.618) ,b = -0.0028 (-0.003272, -0.002413) ,c = 0.13 (0.1085, 0.1525) ,d = -0.000183 (-0.000253, -0.0001131) ; LCPs' BFPs: a = 0.53 (0.4595, 0.601) ,b = -0.0016 (-0.001798, -0.001371),c = 0.083 (0.06994, 0.09596) ,d = -7.5e-05 (-0.0001052, -4.61e-05), **b** *P. reichnowi* : nLCPs' BFPs : a = 0.58 (0.5122, 0.6502), b = -0.0030 (-0.003502, -0.002601) ,c = 0.14 (0.1202, 0.1643),d = -0.000229 (-0.0002978, -0.0001602) ; LCPs' BFPs: a = 0.51 (0.4451, 0.5775), b = -0.0015 (-0.001708, -0.001303), c = 0.08 (0.06701, 0.09343), d = -7.12e-05 (-0.0001015, -4.094e-05). **c** *P. praefalciparum* :nLCPs' BFPs: a = 0.57 (0.5096, 0.6328), b = -0.0030 (-0.003265, -0.002493), c = 0.13 (0.1118, 0.149), d = -0.00018 (-0.0002396, -0.000124); LCPs' BFPs: a = 0.51 (0.4397, 0.5836), b = -0.0016 (-0.00181, -0.001351), c = 0.084 (0.07013, 0.09811), d = -8.149e-05 (-0.0001133, -4.968e-05). **d** *P. adleri* : nLCPs' BFPs : a = 0.14 (0.1177, 0.1604), b = -0.00017 (-0.0002284, -0.0001093), c = 0.50 (0.4461, 0.5571), d = -0.0027 (-0.003079, -0.00226); LCPs' BFPs: a = 0.4437 (0.3759, 0.5114), b = -0.001489 (-0.001727, -0.001252), c = 0.10 (0.09086, 0.1175), d = -6.124e-05 (-8.405e-05, -3.842e-05). **e** *P. gaboni* : nLCPs' BFPs : a = 0.55 (0.4769, 0.6341), b = -0.003121 (-0.003681, -0.00256), c = 0.16 (0.1342, 0.1881, d = -0.0002504 (-0.0003267, -0.0001741); LCPs' BFPs : a = 0.39 (0.3423, 0.4314), b = -0.00124 (-0.001438, -0.001038), c = 0.090 (0.07191, 0.1039), d = -5.822e-05 (-8.891e-05, -2.752e-05). All the coefficients are provided with their 95% confidence bounds

subgenera show distributions located near the center of Pr2 plots, confirming that mutational bias is the predominant factor in shaping codon usage. However, different Pr2 violation patterns are observed. This indicates a differential balance between the influences of mutational bias and natural selection. The LCPs' and nLCPs' centroids of *Vinckeia subgenus* are placed in the second quadrant of the Pr2 plot. This evidence points to a preference for GC-ending codons. The LCPs' centroids of *Laverania* and *Haemamoeba subgenera* are in the first quadrant of their Pr2 plot, respectively. This disposition highlights a preference for AG-ending codons. Differently from *Haemamoeba subgenus*, nLCPs' centroid of *Laverania* parasites appears in the second quadrant of Pr2 plane. This implies, consequently, a preference for C and G in the third codon position. Similar considerations to *Haemamoeba plasmodia* can be drawn for *HIPs*. As shown by LCPs' and nLCPs' centroids, *Simian plasmodia* show a preference for AG-ending codons. Recall the nLCPs of *P. berghei,* that distributed throughout the ENC plane. Similarly, the AT-rich genes of *P. vivax* and *P. cynomolgi,* that are placed on the left side of their ENC planes. These three clusters disappear in their respective Pr2 plots. Therefore, this evidence highlights that 4-fold codon families of these proteins, despite their peculiar ENC distributions, undergo similar selective trends to the other nLCPs. The vectorial nature of Pr2 plots allows for the discrimination of the nucleotide composition of proteins, which can place in the four quadrants. Centroids are therefore insufficient in determining the relative contribution of selective pressure and mutational bias. We, consequently, followed the procedure defined by Forcelloni and Giansanti (2020). We calculated the relative distance (r) of data points in the Pr2 plot from the center, where $A_3 = T_3$ and $G_3 = C_3$, as:

$$r = \sqrt{\left[\frac{A_3}{A_3 + T_3} - 0.5\right]^2 + \left[\frac{G_3}{C_3 + G_3} - 0.5\right]^2}.$$

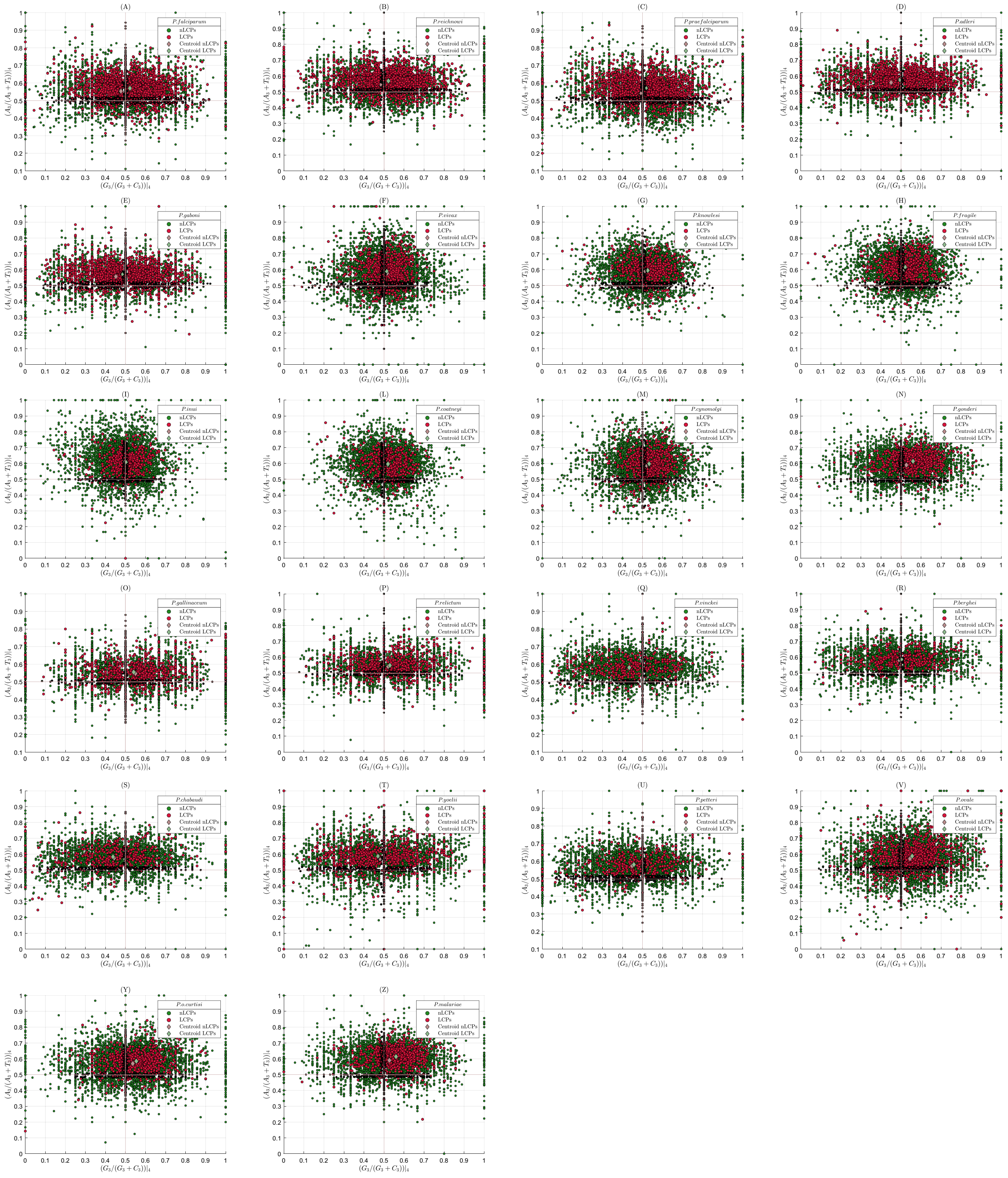

Fig. 10 Pr2 plots. (A, B, C, D, E) Laverania Plasmodia; (F, G, H, I, L, M, N) *Simian plasmodia*; (O, P) *Haemamoeba plasmodia*; (Q,R,S,T,U) *Vinckeia plasmodia*; (V, Y, Z) *Human infectious plasmodia*. Moving away from the Pr2 center returns the extent with which Parity Rule 2 is violated in a Protein Coding Sequence. The more a CDS moves away from the center, the more the contribution to the CUB can be attributed to Selective Pressure. LCPs and nLCPs are represented in red and green, respectively

This distance quantifies the extent of the Pr2 violation pattern associated with the protein-coding sequence under study (Forcelloni and Giansanti 2020). Specifically, $r =$ 0 corresponds to the case of the perfect balance between mutational bias and natural selection. Conversely, if mutational bias and natural selection give different contributions in shaping the CUB of a gene, then its point on the Pr2 plot moves away from the center ($r > 0$). Using this metric, we compared the r-distributions of nLCPs and LCPs (Fig. 11). In line with ENC and SPI results, LCPs are closer to the center than nLCPs (Welch-t-test $p \ll 0.01$ on average, *P. fragile,* $p < 0.01$). Next, we analyzed the extent of the Pr2 violation as a function of the genes' length. In the proposed plan, the Pr2 plot is the abscissa and ordinate plane (x, y), whilst the protein length represents the z-axis. In Fig. 12, we show an example for *Laverania plasmodia* (see Fig. SM.5–8 for the other parasites in SM). We note that, as the protein length increases, the coding sequences get closer to the Pr2 center. This implies that, as the length increases, the Pr2 violation decreases. Alternatively, this implies a more relaxed selective pressure. LCRs are therefore more abundant where Pr2 violation is smaller.

Discussion

In this work, we comparatively studied the proteome of 22 *Plasmodium* species. In this *genus* of successful parasites, we highlighted several aspects of the biology of their low complexity regions. In essence, our work can be summarized by stating that low complexity regions follow a selective and species-dependent evolution as they can be thought of as fingerprints of *Plasmodium* evolution.

Specifically, GC content comparative analysis and correlation between LCR abundance and protein length, as furtherly confirmed throughout the work, highlight that the GC genomic content does not appear as the primary driving force for the stabilization and abundance of LCRs in *Plasmodium* species (Castillo et al. 2019). Moreover, this appears far more evident if we consider the number of LCPs contained in each subgroup (Table SM.1). Even at first glance, indeed, *Laverania subgenus* emerged as a *unicum.*

Nevertheless, globally, *Plasmodium* species can be roughly split into two main groups that are distinguished by their genomic GC content. Nonetheless, the entropy analysis (Shannon 1948), which was furtherly extended in SM (Fig SM. 12–16), provided codon repertoire specific details. This analysis emphasized that, even if *Plasmodia* may have extremely similar genomic GC content, natural selection differently shapes the codon composition of their LCRs.

In line with these first observations, the clustering of LCRs' RSCUs (Fig. 2) corroborates the overall picture. The branching of the dendrogram returned a consistent phylogeny (Tachibana et al. 2012; Valkiunas et al. 2018; Otto et al. 2018; Larson 2019; Kristan et al. 2019; Arisue et al. 2019), highlighting that LCRs are sensitive to the evolutionary

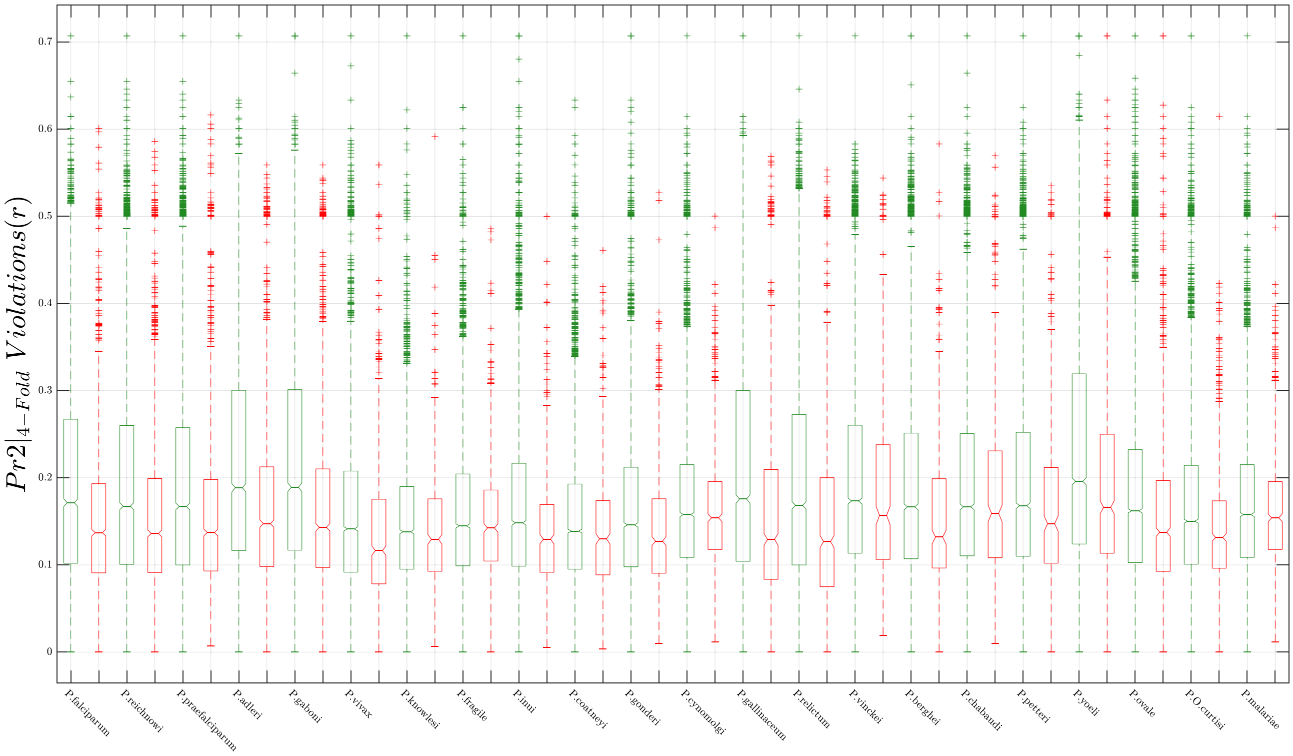

Fig. 11 Pairwise comparison between Pr2 Violation distributions of LCPs and nLCPs. They are represented in red and green, respectively. Medians in each pair differ with 95% confidence (see Mathworks for the explanation of notch function)

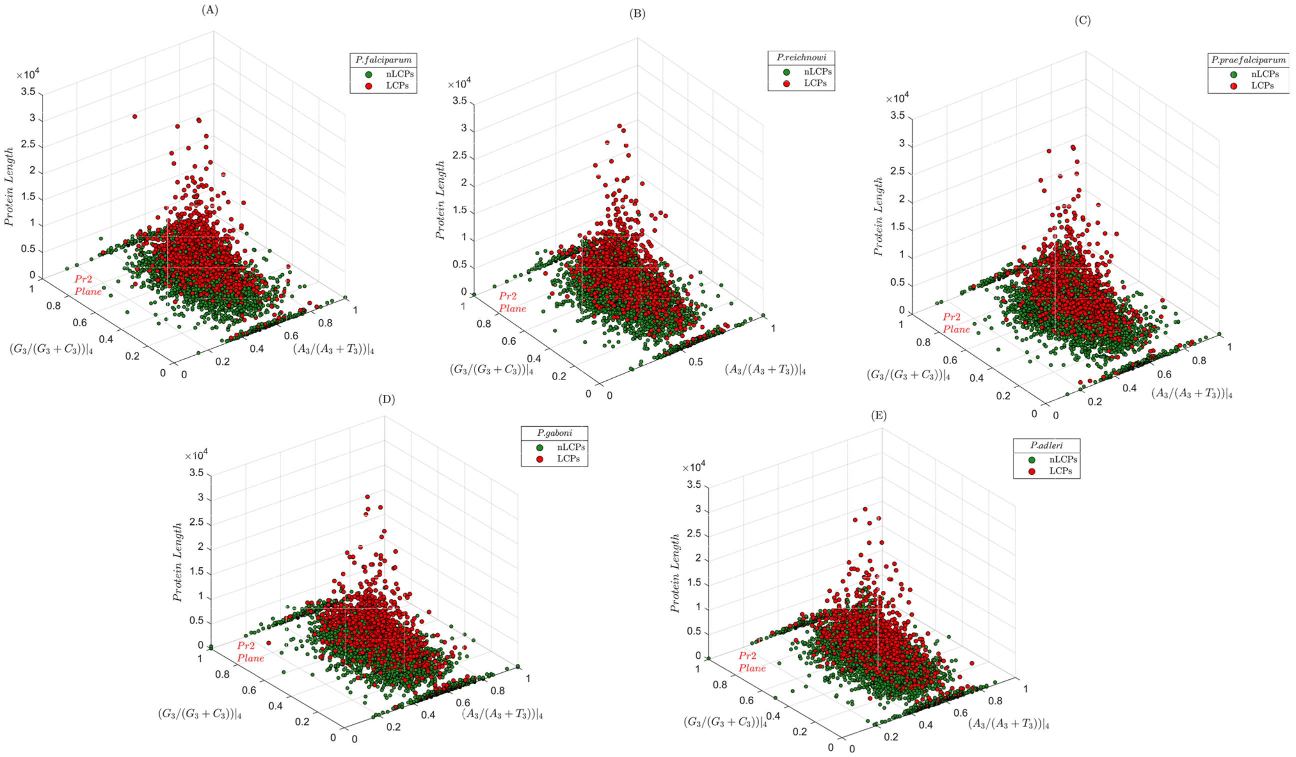

Fig. 12 Correlation between Pr2 data and protein length. Pr2 plot is represented in the (X, Y) plane while protein length (nucleotides) is placed on the z-axis. LCPs and nLCPs are represented in red and green respectively

dynamics of these parasites, suggesting, moreover, LCRs as fingerprints of *Plasmodium* evolution.

The whole-genome analysis based on Effective Number of Codons (Wright 1990; Sun et al. 2013), *Selective Pressure Index,* and Pr2 (Sueoka 1995; Sueoka and Kawanishi 2000), indicated that protein length and a relaxed selective pressure are in tandem with the enrichment in LCRs, typical of LCPs.

Putting our results in perspective, it is worth mentioning that Hamilton and colleagues (2017) performed in vitro tests quantifying the number of transitions and transversions within the genomes of six *P. falciparum* isolates, finally attributing the low genomic GC content of *P. falciparum* to an excess of G: A and C: T transitions. The GC content analyses and the observation of the ENC plots indicate that our results are consistent and corroborate what they observed.

However, we integrate those previous observations, by interpreting the excess of G: A and C: T transitions (Hamilton et al. 2017) as the result of positive selective pressure (not relaxed). Indeed, as the selective pressure decreases, in plasmodial proteins, we noticed a lower violation of the Pr2 rule and an increase in the Effective Number of Codons (SPI and ENC are linearly correlated in each parasite see Table SM.3). From this set of observations, the relaxation of the selective pressure implies a more balanced G: A and C: T ratio within proteins.

Our results cannot sustain the non-adaptive and neutral *conclusions* (our italics) of DePristo et al. (2006). They interpret the abundance of LCRs in *P. falciparum* as a mere consequence of the low GC content of the parasite. ENC, SPI, and Pr2 tests indicate that AT-rich *Plasmodia* undergo similar selective patterns and possess similar levels of guanine and cytosine, or even more extreme as it is for *Haemamoeba subgenus* (Videvall 2018). Our data does not justify the different codon compositions and the pervasiveness of LCRs in them.

Another reason that leads us to infer the presence of Darwinian selection on LCRs in *Plasmodium* species is their amino acid composition. The most frequently chosen codons in all the parasites translate for N, E, D, R, S, and K. These amino acids have similar chemical-physical characteristics such as molecular weight, isoelectric point, and hydrophobicity, and represent, their mutual most common substitutions (NCBI—Amino Acid Explorer). The amino acid selection is therefore conservative (Strachan et al. 2019) as the amino acid substitutions appear selected (evolutionary speaking) preserving LCRs' polypeptide properties. As mentioned above and in agreement with other works (Chaudhry et al. 2018), we retain LCRs to be adaptations of each *Plasmodium* species and not just related, merely, to the GC content of each parasite.

However, LCRs vary in length and amount in proteins (see e.g., Fig. SM.11). Therefore, we do not pronounce ourselves on the phenotype to which LCRs could contribute. A large-scale study concerning their relationship to protein disorder, exploiting recently proposed operational models (Deiana et al. 2019), could shed some light in this regard.

As mentioned, many aspects unite *Plasmodium* parasites. Passing on to protein lengths, indeed, we observed that LCPs are intrinsically longer than nLCPs in each parasite. Previous studies have emphasized that the proteins of *P. falciparum* are longer than the orthologues of other organisms due to the presence of LCRs (Pizzi and Frontali 2001). Other authors (Xue and Forsdyke 2003) report how the LCRs of *P. falciparum* are, as a rule, inserted between protein functional domains, without breaking them. Deferring a more robust verification to future works, we hypothesize that the length of a protein of a parasite is a positive factor for the inclusion of LCRs, because of a dilution effect that reduces the probability of splitting a functional domain. We shall further investigate this interesting point.

Notably, protein lengths and selective pressure were found to be linked. We observed, through SPI and Pr2, that selective pressure decreases with protein length. Consistently, we observed that protein length is positively correlated with the abundance of LCRs, with *Laverania plasmodia* displaying the utmost pronounced tendency to embed these regions within their proteins. Let us stress that, considering SPI and Pr2 correlations with protein length, we can conclude, as one of the main results of this study, that the lower the selective pressure exerted on a plasmodial protein, the more that protein emerges prone to accommodate LCRs.

Trying to interpret this last set of results, it is worth noting that synonymous mutations can have dramatic consequences in cellular processes (Plotkin and Kudla 2011) as they can alter the structure of the mRNA and consequently interfere with the initiation of translation, the stability of the transcript or even with protein folding (Kristofich et al. 2018). Therefore, the most reasonable explanation for these correlations is the same as that provided by Duret and Mouchiroud in their seminal work (1999) for *C. elegans, D. melanogaster*, and *A. thaliana,* and that is ascribable to the dilution effect that we attributed to longer proteins. In fact, we hypothesize that the impact of a broader set of non-optimal codons would be better distributed in a longer nucleotide sequence than in a short one, which intuitively indicates why selection can relax as the protein length increases.

But why the abundance of LCRs is negatively correlated with selective pressure? We propose two answers. The first one is functional innovation. For instance, consider long E-runs that generate distortions in a polypeptide chain (Karlin et al. 2002). A recent report, in particular, has shown in *P. falciparum* that low complexity regions and the glutamate-rich epitopes are highly antigenic, tending to be baits that remove the humoral immunity of the host from the functional domains (Hou et al. 2020). Another example comes from *P. vivax,* where LCRs in PvMSP3α block II seems to guarantee certain phenotypic plasticity (Kebede et al. 2019). LCRs, in this view, could be conceived as a reservoir of chances for natural selection to test new protein functions and adaptations.

The other view concerns the codon purity of LCRs. Especially in AT-rich *Plasmodia*, LCRs are depleted in synonymous codons. Mathematically, the presence of low complexity regions reduces the average number of synonymous codons within a coding sequence, bringing the protein under study closer to a situation of enhanced codon bias. We, then, speculatively suggest that LCRs could act as buffering regions that can modulate, proportionally to their abundance, the overall redundant codon bias of coding sequences. An increased codon bias could modulate in its turn the purifying selective pressure exerted on the protein sequence, as suggested by the observation that non-optimal codons, by their accumulation, can be detrimental for mRNA fitness (Kristofich et al. 2018).

Turning to the technical aspects, we introduced a novel measure of codon bias, SPI, that is based on the ENC plot. Let us briefly comment on the new index. SPI is a way of gauging the Effective Number of Codons (Wright 1990), which we have computed here following the improved method of Sun and colleagues (2013).

The MT1 test has revealed that SPI is particularly useful in statistically evaluating differences or proportions in the codon bias between groups of genes of whole genomes. Compared to *Laverania plasmodia*, *Simian plasmodia* species show, e.g., a lower codon bias (their ENC values are reliably larger, see Fig. 7) which at first sight might suggest that the genomes of *Laverania* are subject to a higher selective pressure. Nevertheless, LPCs of *Simian plasmodia* show off the highest SPI values (see MT1 in results and the interactive MATLAB chart for the Bonferroni Correction in SM), better specifying the contribution of the selective pressure that can be attributed to an ENC value. We, therefore, agree with Gajbhiye and colleagues (2017) in declaring a greater selective pressure on the genes of *P. vivax* compared to those of *P. falciparum*, to which the other members of its family are also added. It is worth noting that SPI is not able to distinguish between positive and negative pressures but is only able to establish what the divergence from Wright's theoretical curve is, measuring, in a normalized scale, how closely the CUB of a gene approximates a situation of extreme bias (20 codons for 20 amino acids).

Let us add here a critical perspective of the choice of the SEG parameters used in this study, a delicate point worth discussing. W = 15 allows the identification of low complexity regions strongly polarized towards certain species of amino acids whilst allowing to find LCRs with a more heterogeneous repertoire of amino acids (Radó-Trilla and Albà 2012). On the other hand, enlarging the window through which SEG is set, introduces a reduction in the LCRs that are identified downstream, where smaller windows (such as W = 6) allow the observation of a larger number of LCRs (Batistuzzi et al. 2016). Comparing the statistics provided by our work with some others (Chaudhry et al. 2018), the identified low complexity regions' amino acids are not starkly different as was also the case for the proportions between nLCPs and LCPs in the parasites that are common to our works. However, our critical sense suggests validating the subjectivity with which SEG is triggered, finding a trade-off between the approach we followed (Radó-Trilla and Albà 2012) and more refined investigations such as those proposed by Batistuzzi and colleagues (2016), identifying any unclassified LCP we did not find.

Likewise, it is important to point out that GC content often aggravates genome assemblies (Chen et al. 2013). Even though the additional tests on protein length and SPI distributions reported in the Supplementary Materials confer certain robustness to our observations (Fig. SM. 9–10) this risk is not absent.

Despite the many points still to be clarified and investigated, this study opens, we believe, a spectrum of methods and concepts to motivate future research on the global biology of parasitism. Indeed, extending the present study to other *Apicomplexa* would improve the current understanding of the evolutionary success of this large class of parasites.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10709-021-00126-6.

Acknowledgements The authors want to thank the anonymous referees whose corrections and questions have allowed us to greatly improve the work. Furthermore, we thank our Editor, prof. Horacio Naveira, for the extraordinary editing service and management of the entire peer review process.

Authors contribution AC and AG conceived the study. AC conducted the analyses and conceived the original manuscript. AC, SF and AG read, edited, and approved the final manuscript.

Data availability Data are retrievable from public archives as mentioned in methods. Supplementary files are provided as electronically.

Code availability Scripts are available, upon request, from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict or competing of interest.

References

Albà MM, Guigó R (2004) Comparative analysis of amino acid repeats in rodents and humans. Genome Res 14:549–554. https://doi.org/10.1101/gr.1925704

Arisue N et al (2019) Apicoplast phylogeny reveals the position of *Plasmodium vivax* basal to the Asian primate malaria parasite clade. Sci Rep 9:7274. https://doi.org/10.1038/s41598-019-43831-1

Battistuzzi FU et al (2016) Profiles of low complexity regions in Apicomplexa. BMC Evolut Biol 16:47. https://doi.org/10.1186/s12862-016-0625-0

Carter R, Walliker D (1975) New observations on the malaria parasites of rodents of the Central African Republic - Plasmodium vinckei petteri subsp. nov. and Plasmodium chabaudi Landau, 1965. Ann Trop Med Parasitol 69:187–196. https://doi.org/10.1080/00034983.1975.11687000

Castillo AI, Nelson A, Lyons E (2019) Tail wags the dog? Functional gene classes driving genome-wide gc content in plasmodium spp. Genome Biol Evol 11:497–507. https://doi.org/10.1093/gbe/evz015

Chaudhry SR et al (2018) Comparative analysis of low complexity regions in Plasmodia. Sci Rep 8:335. https://doi.org/10.1038/s41598-017-18695-y

Chen YC et al (2013) Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One. https://doi.org/10.1371/journal.pone.0062856

Collins WE, Jeffery GM (2005) Plasmodium ovale: parasite and disease. Clinical microbiology reviews 18:570–581. https://doi.org/10.1128/CMR.18.3.570-581.2005

Collins WE, Jeffery GM (2007) Plasmodium malariae: parasite and disease. Clin Microbiol Rev 20:579–592. https://doi.org/10.1128/CMR.00027-07

Corradetti A, Garnham PCC, Laird M (1963) New classification of the avian malaria parasites. Parassitologia 5:1–4

Deiana A et al (2019) Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. PLoS One. https://doi.org/10.1371/journal.pone.0217889

DePristo MA, Zilversmit MM, Hartl DL (2006) On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene 378:19–30. https://doi.org/10.1016/j.gene.2006.03.023

Duret L, Mouchiroud D (1999) Expression pattern and surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceed Natl Acad Sci USA 96:4482–4487. https://doi.org/10.1073/pnas.96.8.4482

Escalante AA, Ayala FJ (1994) Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proceed Natl Acad Sci USA 91:11373–11377. https://doi.org/10.1073/pnas.91.24.11373

Faux NG et al (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 15:537–551. https://doi.org/10.1101/gr.3096505

Ferreira et al (2003) Sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-1 (MSP-1) of Plasmodium falciparum. Gene 304:65–75. https://doi.org/10.1016/s0378-1119(02)01180-0

Filisetti et al (2013) Aminoacylation of Plasmodium falciparum tRNA (Asn) and insights in the synthesis of asparagine repeats. J Biol Chem 288:36361–36371. https://doi.org/10.1074/jbc.M113.522896

Forcelloni S, Giansanti A (2020) Evolutionary forces and codon bias in different flavors of intrinsic disorder in the human proteome. J Mol Evol 88:164–178. https://doi.org/10.1007/s00239-019-09921-4

Forsdyke D (2016) Evolutionary Bioinformatics. Springer

Frugier M et al (2010) Low complexity regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. FEBS Letters 584:448–454. https://doi.org/10.1016/j.febslet.2009.11.004

Gajbhiye S, Patra PK, Yadav MK (2017) New insights into the factors affecting synonymous codon usage in human infecting Plasmodium species. Acta Trop 176:29–33. https://doi.org/10.1016/j.actatropica.2017.07.025

Garnham PC (1964) The subgenera of plasmodium in mammals. Ann Soc Belges Med Trop Parasitol Mycol 44:267–271

Gemayel R et al (2012) Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes 3:461–480. https://doi.org/10.3390/genes3030461

Gragg H, Harfe BD, Jinks-Robertson S (2002) Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in Saccharomyces cerevisiae. Mol Cel Biol 22:8756–8762. https://doi.org/10.1128/mcb.22.24.8756-8762.2002

Hamilton WL et al (2017) Extreme mutation bias and high AT content in Plasmodium falciparum. Nucleic Acids Res 45:1889–1901. https://doi.org/10.1093/nar/gkw1259

Hou N et al (2020) Low-complexity repetitive epitopes of plasmodium falciparum are decoys for humoural immune responses. Front Immunol 11:610. https://doi.org/10.3389/fimmu.2020.00610

Howes RE et al (2016) Global epidemiology of Plasmodium vivax. Am J Trop Med Hygiene 95:15–34. https://doi.org/10.4269/ajtmh.16-0141

Karlin S et al (2002) Amino acid runs in eukaryotic proteomes and disease associations. Proceed Natl Acad Sci USA 99:333–338. https://doi.org/10.1073/pnas.012608599

Kebede AM et al (2019) Effect of low complexity regions within the PvMSP3α block II on the tertiary structure of the protein and implications to immune escape mechanisms. BMC Str Biol 19:6. https://doi.org/10.1186/s12900-019-0104-0

Kristan M et al (2019) Mosquito and human hepatocyte infections with *Plasmodium ovale curtisi* and *Plasmodium ovale wallikeri*. Trans Royal Soc Trop Med Hygiene 113:617–622. https://doi.org/10.1093/trstmh/trz048

Kristofich J et al (2018) Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. PLoS Genet. https://doi.org/10.1371/journal.pgen.1007615

Larson B (2019) Origin of Two Most Virulent Agents of Human Malaria: *Plasmodium falciparum* and *Plasmodium* vivax. https://doi.org/10.5772/intechopen.84481

Legendre M et al (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17:1787–1796. https://doi.org/10.1101/gr.6554007

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221. https://doi.org/10.1093/oxfordjournals.molbev.a040442

Muralidharan V, Goldberg DE (2013) Asparagine repeats in *Plasmodium falciparum* proteins: good for nothing? PLoS Pathogens. https://doi.org/10.1371/journal.ppat.1003488

Muralidharan V et al (2012) Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers. Nat Commun 3:1310. https://doi.org/10.1038/ncomms2306

Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394. https://doi.org/10.1093/oxfordjournals.molbev.a004201

Otto TD et al (2018) Genomes of all known members of a *Plasmodium subgenus* reveal paths to virulent human malaria. Nat Microbiol 3:687–697. https://doi.org/10.1038/s41564-018-0162-2

Pizzi E, Frontali C (2001) Low-complexity regions in *Plasmodium falciparum* proteins. Genome Res 11:218–229. https://doi.org/10.1101/gr.gr-1522r

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev 12:32–42. https://doi.org/10.1038/nrg2899

Prugnolle F et al (2010) African great apes are natural hosts of multiple related malaria species, including *Plasmodium falciparum*. Proceed Natl Acad Sci USA 107:1458–1463. https://doi.org/10.1073/pnas.0914440107

Radó-Trilla N, Albà M (2012) Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evolut Biol 12:155. https://doi.org/10.1186/1471-2148-12-155

Rich SM, Xu G (2011) Resolving the phylogeny of malaria parasites. Proceed Natl Acad Sci USA 108:12973–12974. https://doi.org/10.1073/pnas.1110141108

Saitou N (2018) Introduction to evolutionary genomics. Springer. https://doi.org/10.1007/978-1-4471-5304-7

Salichs E et al (2009) Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. PLoS Genet. https://doi.org/10.1371/journal.pgen.1000397

Sato S (2021) *Plasmodium*-a brief introduction to the parasites causing human malaria and their basic biology. J Physiol Anthropol 40:1. https://doi.org/10.1186/s40101-020-00251-9

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.hdl:10338.dmlcz/101429

Sharp PM, Li WH (1987) The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15:1281–1295. https://doi.org/10.1093/nar/15.3.1281

Shen H, Kan JL, Green MR (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. Molecular Cell 13:367–376. https://doi.org/10.1016/s1097-2765(04)00025-5

Strachan T, Goodship J, Chinnery P (2019) Genetica e genomica nelle scienze mediche. Zanichelli

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325. https://doi.org/10.1007/BF00163236

Sueoka N, Kawanishi Y (2000) DNA G+C content of the third codon position and codon usage biases of human genes. Gene 261:53–62. https://doi.org/10.1016/s0378-1119(00)00480-7

Sun X, Yang Q, Xia X (2013) An improved implementation of effective number of codons (nc). Mol Biol Evol 30:191–196. https://doi.org/10.1093/molbev/mss201

Tachibana S et al (2012) Plasmodium cynomolgi genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat Genet 44:1051–1055. https://doi.org/10.1038/ng.2375

Toll-Riera M et al (2012) Role of low-complexity sequences in the formation of novel protein coding sequences. Mol Biol ution 29:883–886. https://doi.org/10.1093/molbev/msr263

Valkiunas G (2000) Avian malaria parasites and other haemosporidia. CRC Press, NW Corporate Blvd., Boca Raton, Florida

Valkiunas G et al (2018) Characterization of Plasmodium relictum, a cosmopolitan agent of avian malaria. https://doi.org/10.1186/s12936-018-2325-2

Verstrepen KJ et al (2005) Intragenic tandem repeats generate functional variability. Nature Genet 37:986–990. https://doi.org/10.1038/ng1618

Videvall E (2018) Plasmodium parasites of birds have the most AT-rich genes of eukaryotes. Microbial Genom. https://doi.org/10.1099/mgen.0.000150

Waters AP, Higgins DG, McCutchan TF (1993) Evolutionary relatedness of some primate models of Plasmodium. Mol Biol Evol 10:914–923. https://doi.org/10.1093/oxfordjournals.molbev.a040038

Waters AP, Higgins DG, McCutchan TF (1993) The phylogeny of malaria: a useful study. Parasitol Today (Personal ed.) 9:246–250. https://doi.org/10.1016/0169-4758(93)90066-o

Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–571. https://doi.org/10.1016/s0076-6879(96)66035-2

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Xue HY, Forsdyke DR (2003) Low-complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. Mol Biochem Parasitol 128:21–32. https://doi.org/10.1016/s0166-6851(03)00039-2

Yadav MK, Swati D (2012) Comparative genome analysis of six malarial parasites using codon usage bias based tools. Bioinformation 8:1230–1239. https://doi.org/10.6026/97320630081230