Title page

Article title:
Adaptive and Conservative Protein Assets In Plasmodia. Two Different Scenarios of the Biology of Malaria Parasites

Authors: 1*Andrea Cappannini, 1* Sergio Forcelloni and 1, 2 Andrea Giansanti

Affiliation:
1 Sapienza, University of Rome, Department of Physics, P.le A. Moro 5, 00185 Roma, Italy.
2 Istituto Nazionale di Fisica Nucleare, INFN, Roma1 section. 00185, Roma, Italy.

Corresponding Author:
Andrea Cappannini
Sapienza University of Rome, Department of Physics, P.le A. Moro 5, 00185 Roma, Italy.
E-mail: *andreacappannini@gmail.com*
Phone: *+39 3493365228*

Abstract

The proteins of Plasmodia are rich in low complexity traits, that should be involved in the evolutionary history of this class of parasites. Considering the proteomes of 22 plasmodia we investigate here how the length of the proteins, the Low Complexity content, the evolutionary pressure and other biological features are correlated. From our data a scenario emerges in which the incorporation of low complexity traits appears as a feature of long proteins of relaxed purifying pressure and a potential source of functional innovation. A scenario able to capture the evolutive success of Plasmodia as parasites.

Introduction

Low Complexity regions (LCRs) or non-random regions of few amino acids are ubiquitously present in proteins (Kumari et al. 2015). Despite their evolutionary history and functional roles remain unclear (Batistuzzi et al. 2016) they are evolutionary conserved being present in all the kingdom of life from Bacteria to, Archaea to Eukarya (Wootton & Federhen, 1996).The replication slippage (RS), a mutational process that expands or contracts microsatellites, *i.e.*, repetitive genetic sequences in which repeating units contain from one to six bases, is proposed as the major factor that determines the genetic origin of LCRs (Guy-Franck et al. 2008;Mosbach et al. 2019;Ellegren, 2004). The expansion or contraction of these regions can greatly diversify the space of protein sequences, generating novel cellular functions with minimal genetic load (Kashi & King, 2006; Toll-Riera et al. 2011) suggesting a crucial evolutive and adapting implication. Despite being a prolific source of quantitative and qualitative variation (Kashi & King, 2006), LCRs are frequently involved in neurodegenerative (Gatchel & Zoghbi, 2005) and developmental diseases (Brown & Brown, 2004), gaining or losing protein functions or altering protein conformations ultimately leading to changes in protein activity or abundance (Gatchel & Zoghbi, 2005). On the other hand, they furtherly play a central role in protein-protein interactions (Shen et al. 2004) and protein sub-cellular localization (Salichs et al. 2009). *P.falciparum*, the etiological agent of the most virulent form of human malaria (Pizzi & Frontali, 2001; Otto et al. 2018) shows a vast content of Low Complexity (LC) residues strongly enriched in asparagine (N) (Pizzi & Frontali, 2001;Filisetti et al. 2013;Frugier et al.2009). As far as their function is concerned, it was proposed they can behave as an immune-evasive mechanism (Pizzi & Frontali, 2001) or as tRNA sponges (Filisetti et al. 2013; Frugier et al.2009). However, different lines of evidence bring to consider a different role for these regions. Asparagine has been found prion forming after heat shocks. This could happen due to the infection, after the occurrence of the febrile status of the host. This would lead to retain N residues as a potential drawback for the parasite. However, N residues have been found in reducing proteo-toxicity unlike others prion forming residues (Halfmann et al. 2011). In an extremely interesting fashion, heat shock proteins, in particular the Exported Chaperone *Pf*Hsp110c, prevent this aberrant aggregation (Muralidharan et al. 2012; Muralidharan & Golberg, 2013). Muralidharan and Golberg (2013) proposed that with the feedback of heat shock proteins, N repetitive stretches would remain unfolded collecting mutations and gradually leading the protein structures to gain new folds and functions. Therefore, asparagine seems to represent a certain kind of benefit for *P. falciparum* appearing as the result of a positive selection or luckily, an evolutionary accident (Muralidharan & Golberg, 2013). Other results sustain that N-stretches are the outcomes also of the selection (Pizzi & Frontali, 2001) retaining that the extreme Adenine richness could not be the unique gauge up to determine their genetic origin, while others stresses the selection of an initial codon seed replicated by the action of replication slippage (Chaudhry et al. 2018). Paying attention to the genomic level other results enlighten how proteins have to adapt to the demands of the environment, but also to pressures that have derived not from *conventional phenotype*, but from other environments including the "*internal reproductive environment*" acting on the genome phenotype (Forsdyke, 2016;Bernardi & Bernardi,1986a). Thus, these strands could have been inserted to respond to the needs of the *P.falciparum* in order to extremize its AT content as it has been found in several works (Pizzi & Frontali, 2001; Filisetti et al. 2013;Frugier et al.2009; Muralidharan & Golberg, 2013; Chaudhry et al. 2018). Since nucleotide content cannot be extremized without affecting the phenotype of a protein these strands would have been inserted initially as mere *placeholders* to further enhance the AT content and then acquiring or not a classical phenotype importance (Forsdyke, 2016). As far as our results are concerned, we analyzed different facets of the biology of this genus considering 22 species among which *P. falciparum*. To have a

reference frame, we divided each proteome in two categories i.e. LCPs, Low Complexity Containing Proteins, namely Proteins that contain at least one LCR inside their chain and not Low Complexity Containing Proteins (nLCPs), their counterpart with no LC content. As seen in many works, protein length has been subject of many studies. A strong correlation between LCR insertions and protein length was first spotted in *P. falciparum* by Pizzi and Frontali (2001). Others instead, investigated the implication of genome pressures with the addition of LCRs in some biological entities (*e.g. Epstein bar virus*) among which *P. falciparum* (Forsdyke, 2016). Furthermore, protein length resulted to be an interesting metric to compare the evolutionary age of proteins, where shorter ones appear to have emerged in ancient times (Lipman et al. 2002). On the other hand, longer proteins appear to be innovative from an evolutionary point of view (Wang et al. 2005). On this aim, we decided to further extend this context related knowledge to the other plasmodia. We find that the rate of insertion is not as strong as it is for *P. falciparum* and *Laverania subgenus* where LCRs are inserted with strength. This suggests that the whimsical LC content is a trait of these species. Furthermore, we find that LCPs result to be longer than nLCPs regardless of their LC content in each organism. This suggests that the protein length is one of the causative characteristics that drive the insertion of Low Complexity segments (LCSs). In addition, observing the amino acid composition of LCRs we further sustain that the overabundance of LC-asparagine is special trait of *Laverania*, maybe derived from *Haemamoeba* parasites that dispose a similar N-richness. To get other important evolutionary insights, we implemented The Effective Number of Codons (ENC, Wright,1990) by its improved version proposed by Sun and colleagues (2013). Our results clearly show two scenarios: nLCPs appear to be closed in a more conservative shell where the purifying pressure restrains their CUB suggesting a crucial role in maintaining the fitness of these organisms. On the other hand, the CUB found in LCPs is more redundant, indicating that the mutational bias acts as the priority body. ENC is affected by redundancies that impede to clearly understand the extent of the action of Selective and mutational pressure. Proposing a new index that here we call the SPI, we strengthen our former observation. We show that LCPs depart the less from the theoretical curve with respect to nLCPs. In turn, this implies that their CUB departs the less from the null model of no-codon preference. Overall, we propose LCPs as recently evolved, LCRs as secondary DNA structures that prevent the mutagenic process to impair the most important protein regions and nLCPs as potential knock-down targets to strike the malaria parasites worldwide.

2 Methods

2.1 Data Sources and Filtering

The whole set of coding DNA sequences (CDS) for each plasmodia was retrieved from PlasmoDB (https://plasmodb.org) and NCBI GenBank (ftp://ftp.ncbi.nih.gov), respectively. We considered only CDSs that begin with the start codon (AUG), ending with a stop codon (UAG, UAA, or UGA), and having a multiple length of three. Incomplete and duplicated genes, with internal gaps or not recognized nucleotides were excluded. Amino-acid sequences were translated by *nt2aa* MATLAB function.

2.2 Prediction of Low Complexity regions

LCRs were retrieved using SEG algorithm with standard parameters (Wootton & Federhen, 1996).

2.3 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. Similarly, it is possible to define the GC-content in the first (GC1), second (GC2), and third (GC3) codon positions, as follows:

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

2.4 ENC, Effective Number of Codons

We calculated the effective number of codons (ENC) to estimate the extent of the codon usage bias of plasmodium genes. The values of ENC range from 20 (when just one codon is used for each amino acid) to 61 (when all synonymous codons are equally used for each amino acid) (Wright, 1990). Thus, the smaller is the ENC value the larger is the extent of codon preference in a gene. The ENC values of all individual genes (>300 bp) were calculated using the ENC improved implementation proposed by Sun et al. (2013). The codon table is separated in synonymous codon families where the 6-fold codon families (Leu, Ser, and Arg) were divided into 2-fold and 4-fold codon families since these two subsets can be under a dramatic different selection (Sun et al. 2013). For each coding sequence, we quantify Fα, 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 codon *i* of the codon family α in the coding sequence, and $n_\alpha = \sum_{k=1}^{m_\alpha} n_{k_\alpha}$. Finally, the gene - specific ENC is defined as:

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

where N_s is the number of codon families with a single codon (i.e., Met and Trp codon families) and K_m is the number of families with degeneracy m. The ENC-plot is a plot in which the ENC is the ordinate and the GC3 is the abscissa (Forcelloni & Giansanti, 2019). 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 GC3:

$$ENC = 2 + GC_3 + \frac{29}{GC_3^2 + (1 - GC_3^2)^2}$$

Genes for which the codon choice is only constrained by mutational bias are expected to lie on or just below the Wright's theoretical curve (WTC). So, points that fall below the bell-shaped line suggest deviations from a null model of no-codon preference (Novembre, 2002). In this case, the vertical distance between the point and the theoretical curve provides an estimation of the relative extent to which natural selection and mutational bias affect CUB.

2.5 Selective Pressure Index (SPI)

The vertical distance between a point and the WTC provides an estimation of the relative extent to which natural selection and mutational bias affect the codon usage in genes (Novembre, 2002). Looking at the vertical distance that divide a protein from the null model of no-codon preference represented by the WTC, it

can be observed that the inclusion or the exclusion of one codon can contribute in a very different fashion to the deviation from the WTC. In fact, the same ENC value can represent proteins under a more relaxed or a more stringent purifying pressure. Therefore, depending on the GC_3 content the inclusion or the extrusion of a codon from the range of the possible synonyms that can be chosen to convey the genetic information, can represent a more stringent or a more lenient action of the natural selection in the general optimization process that can lead the coding sequence toward the situation of extreme bias (20 codons for 20 amino acids) or to its extreme contrary. We thus propose the SPI that is formally defined as follows:

$$SPI(GC_3) = \frac{ENC_w(GC_3) - ENC_p(GC_3)}{ENC_w(GC_3) - 20}$$

ENCp is the ENC score for the CDS under study that will be characterized by a particular GC_3 content; ENCw is the ENC value of the WTC for that GC_3 value. This distance is normalized over the distance from the extreme bias situation for that GC_3value. SPI helps to compare the distances under the WTC being a normalized index in the range of [0, 1]. Then, if SPI = 0 the gene will fall on the WTC (i.e., no selection). On the other hand, if SPI = 1 then the CUB of the gene will be subject to a strong selection (situation of extreme bias) where 20 sense codons will be used for 20 amino acids.

2.5 LCPs and nLCPs

To have a reference frame to compare evolutionary pressures we divided the various CDS sets in two branches, i.e. LCPs, with at least one LCR inside their chain and nLCPs that have none. We translated the CDS into amino acids, identifying the LCRs. Therefore, we indexed these proteins taking the corresponding nucleotide chains conserved in the original FASTA file.

2.6 Cladistic and Organism Overview

The distant relation of *P. falciparum* with other human malaria parasites permitted to put and separate this organism in a distinct *monophyletic subgenus* termed *Laverania* (Sibley et al.2012; Liu et al. 2016). In 2009 the only other known laverania plasmodium was *P.reichnowi* (Otto et al. 2018). At least six other species was added with three chimpanzees infecting plasmodia (*P.reichnowi, P.billcollinsi and P.gaboni*) and three that infect gorillas *(P.praefalciparum, P.blacklocki* and *P.alderi*).Given their availability, we retrieved the CDS sets for *P.falciparum, P.praefalciparum, P.reichnowi, P.alderi, P.gaboni*. Among *Laverania, P.falciparum* is the only that successfully infects humans (Otto et al. 2018) but despite being considered for a long time strictly human infecting, it has been found also in gorillas (Prugnolle et al. 2010) raising concerns about possible reciprocal host transfer of these parasites. Passing from these African parasites (Otto et al. 2018) to other geographical species, the radiation that led to the evolutionary bush of Asian monkey parasites is thought to be a recent event with diverse evolutive traits appeared *de novo* (Escalante et al. 2019). Among these, it is found *P.vivax*. While *P.falciparum* it's the deadliest among the *genus of plasmodium* (Pizzi & Frontali, 2001;Otto et al. 2018), given its morbidity, *P.vivax* places itself at the second place of the podium, due to its extremely wide geographical coverage (Howes et al. 2016) putting 2.5 billion people, from Asia to South America (Vogel,2013), at risk of infection. *P.vivax* is evolutionary related to other species of plasmodia naturally infecting Asian primates (Escalante et al. 1998) showing morphological and biological

characteristics with several macaque infecting plasmodia. Among these *P.knowlesi* has recently emerged as a human pathogen (Rich & Xu, 2011). Thought to be strictly a primate infecting plasmodium, naturally acquired human infections for *P.knowlesi* were thought rare (Faust & Dobson, 2015) until a large focus of human contagions in 2004 in Sarawak, Malaysian Borneo (Singh et al., 2013). In this subgroup we analyzed also *P.cynomolgi* and *P.coatney* (Galinski & Barnwell, 2012), *P.inui*, that is found to be related to *P.cynomolgi , P.coatney and P.fragile*, a naturally bonnet macaques infecting plasmodium (Galland, 2000). Its phylogenetic relationship comprises also *P.vivax* (Mitsui et al. 2010). We considered also *P.gonderi* that, despite being a naturally infecting *cercocebus albigena* plasmodium (a primate from Senegal), has been found related to *P.vivax* depicting a possible African origin for these parasites (Arisue et al. 2019). We refer to this set of parasites as *Simian Plasmodia*. Furtherly, to have a wider vision about the evolutionary and quantitative diversification of LCRs, and in general of the plasmodias' protein assets, we considered *rodents plasmodia* (*Vinckeia subgenus*). Among these we were able to retrieve complete CDS sets for *P.vinckei* (Carter & Wallinker, 1975) and other subspecies, i.e. *P.petteri, P.berghei and P.yoeli* (Garnham, 1964). Two specimens of the *Haemamaeba subgenus* (Corradetti et al. 1963) were considered: *P.gallinaceum* and *P.relictum*. Several lines of evidence address the origin of *P.falciparum* proposing avian plasmodia as possible ancestors (Waters et al. 1991; Waters et al. 1993 a-b;McCutchan et al. 1996;Escalante & Ayala,1994). This evolutionary radiation luckily appeared as the result of a lateral transfer between human and avian host (Waters et al. 1991), probably when avian were originally domesticated (Escalante et al. 1998). Eventually we considered *P.relictum* that is one of the most geographically widespread malaria parasite for birds (Valkiunas, 2005). At last, we considered other human infecting plasmodia (OHIP), i.e. *P.ovale* causing tertian malaria (Collins & Jeffery, 2005), *P.ovale curtisi* in a strict evolutionary relation with the former one (Kristan et al. 2019) and *P.malariae* causing quartan malaria (Collins & Jeffery, 2007). We want to emphasize that establishing the various phylogenetic relationships is beyond our scope. However, we found it necessary to give a brief description of these organisms defining the criteria and the motivations by which they were placed in a sub-group rather than in another.

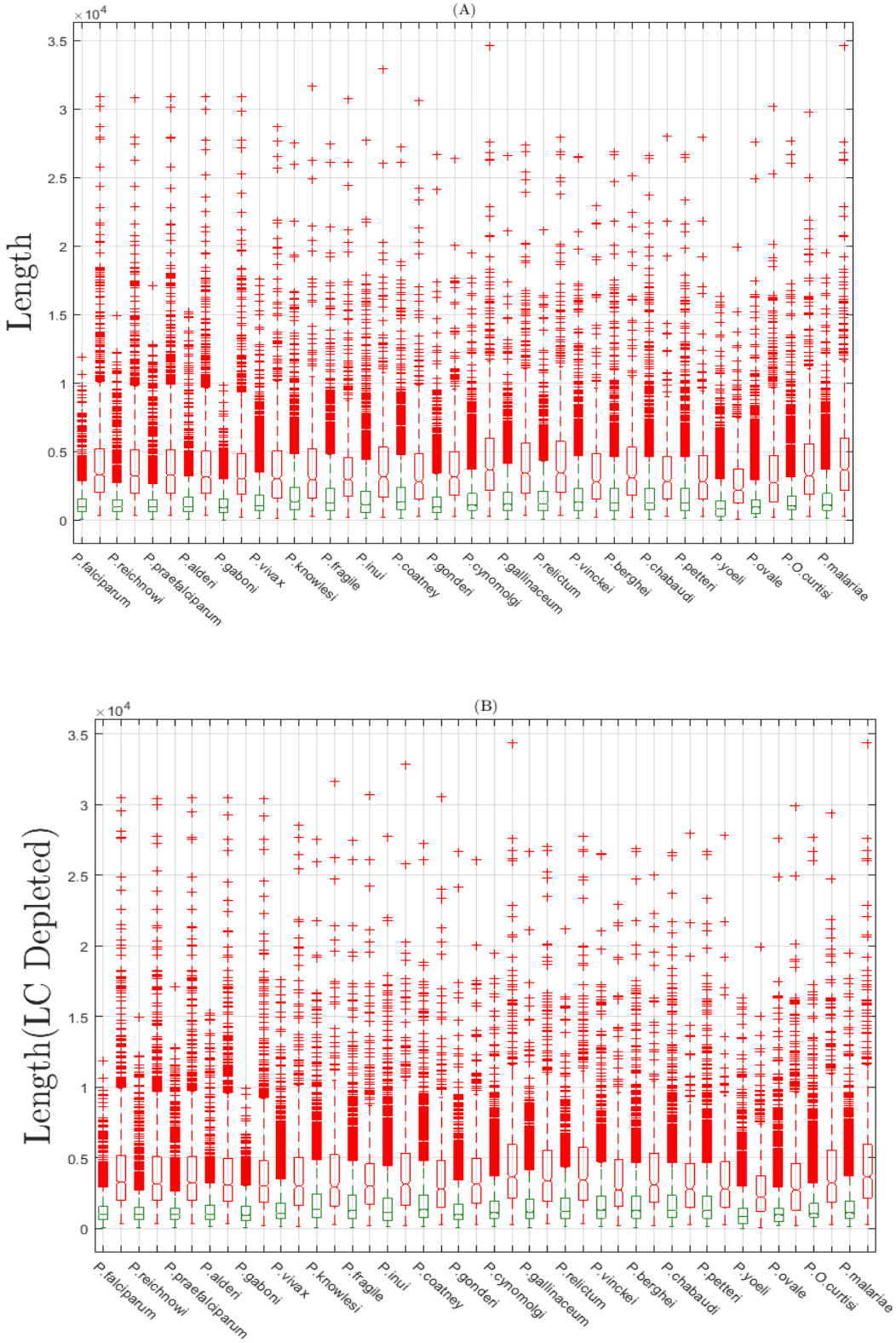


Fig 1 *Pairwise comparison for each organims for LCP(red) and nLCP(green) amino acid protein length distributions. Fig 1 (A) Protein length for LCPs is calculated with low complexity inside the protein chain. (B) Protein length for LCPs is calculted with the depletion of LCRs. Each pairwise comparison is significant (p<<0,00001)*

3 Results

3.1 Protein Length, GC and LC content

Protein length is generally linked to protein function with the wide variance of protein lengths in an organism being representative of a vast diversity of biological functions and roles (Lipman et al. 2002). However, it has to be considered that the coding sequence behind the amino-acid sequence is the result of a compromise between the needs of DNA function, RNA function and protein function (Pancsa & Tompa, 2016) putting the *classical phenotype* needs among the other necessities related to the internal reproductive environment (Forsdyke, 2016;Bernardi &Bernardi,1986a). To have more structural and evolutionary insights of the biology of *P. falciparum* and its LCRs, we investigated over all the organisms at our disposal, comparing LCPs' and nLCPs' amino acid protein length distributions. On this aim, we implemented a pairwise comparison juxtaposing each organism in its nLCPs' and LCPs' length distributions. In Fig.1(A) we represented LCPs and nLCPs, respectively in red and green. We find that LCPs are significantly longer than nLCPs (*Welch t-test*; $p<<0.01$) in each organism. We further implemented the *MATLAB* notch function. It permits to observe that when two notches do not overlap, medians do differ with 95% confidence (Mathworks).We observe that the medians between each LCP-nLCP pair differ with significance. Pizzi & Frontali (2001) observed how *P. falciparum* proteins were longer than the orthologs found in other organisms leading to the conclusion that LCRs do not impair protein functions. Consequently, despite the evident structural difference that sunder nLCPs and LCPs, we wondered if the insertion rate of the LCSs occurred in the same way in each parasite. Furthermore, we wondered if the length of the LCPs was a direct consequence of the presence of the LCRs or if it was a feature *de facto*. In the first instance, protein length would have been a favorability index for the insertion of LCRs. In the second instance, LCRs would be causative of LCPs' protein length. On this aim, we implemented a set of correlation analysis (*Pearson*) between each protein length and its LC content in each parasite. Tab.1 collects these results together with other statistics. The first column of Tab.1 disposes all the correlation coefficients (all the correlations were significant $p<<0,01$). *Laverania* subgenus flaunts the highest correlation coefficients. It follows that the longer the protein the higher the probability for LCRs to squeeze into interdomains. The same strength of insertion is not observed in all the remaining parasites. At first glance, this suggested that, regardless of the LC content, LCPs are longer than nLCPs. To further clarify the conundrum, we implemented another pairwise comparison depleting protein chains from their LC content. Results are represented in Fig 1 (B). The pairwise comparison further confirms our precedent hypothesis, namely, LCPs appear to be longer than nLCPs *de facto* (each pairwise comparison is significant *welch t-test* $p << 0,01$). Therefore, protein length results to be a favorability index for the LCRs insertion. We pondered about the relation between GC and LC content. We observe a quint LC content in *Laverania*. To further clarify the role of the selection and of the GC content we made a series of comparisons, first juxtaposing in pair each group (*Mann Whitney tests*) considering the LC content distribution of each subgenus. *P. malariae* was cut off given no specific phylogenetic relationship with the other parasites (Collins & Jeffery, 2007). We find that LC content in *Laverania* differ with significance from *Vinckeia* and *Simian plasmodia* ($p<<0,001$). No significative differences are found with *Haemamoeba* and the subgroup composed by *P.ovale* and *P.o.curtisi* ($p>0,05$). The second column of Tab. 1 collects the average GC content of each parasite. It is possible to note a symmetry in LC-GC content between *Laverania subgenus* and *simian plasmodia*. *Laverania* show at the same time a massive content of LC and a low GC %. Contrarily, *Simian plasmodia* are characterized by a poorer content of LC residues and a less biased GC %. De facto, the former subgroup is found to be extremely AT rich with the highest number of LC residues and the highest correlation coefficients. Symmetrically, the latter shows a more averaged GC content with a drastic reduction in LC residues and correlations. This suggests a certain impact of the GC content on the elongation of LCRs. However, looking at the other subgroups (*e.g. Vinckeia*) we observe a puzzling situation. *Vinckeia* appear to be close to *P. falciparum* in term of codon usage (Yadav & Swati, 2012) reflecting a certain phylogenetic link already stressed in other works (Waters et al. 1993). However, we do not find the same massive expression of LC residues in face of the low GC content (LC content distributions in the two *subgenus* differ with significance, $p < 0,01$). Moreover, we observe in *Vinckeia* low correlations between protein length and low complexity. It

appears reinforcing to the observation made by Pizzi & Frontali (2001) who state that in *P. falciparum* A richness could not be the unique driving force behind the presence of LC. This implies that the explosion of Low Complexity could be a distinctive trait of *Laverania subgenus.* Conversely, *Haemamoeba* parasites see instead a closer behavior to *Laverania* (their LC distribution does not differ with significance, $p<0,05$). However, the lack of specimens and the statistical pauperism make it necessary to have more datasets to fortify and clarify these assumptions. Eventually, we find *P.ovale* and *P.o.curtisi* having a GC content halfway between *simian plasmodia* and *Laverania* with a rather high LC content (their LC distribution does not differ with significance from *Laverania*, $p>0,05$). The same goes for *P. malariae* showing similar features with *P.ovale* and *P.o.curtisi.* Overall, GC content alone appear not to be the unique feature that shapes Low Complexity Regions. Further analysis by the observation of other genome pressures (*e.g. purine-pyrimidine loading pressures*) can help to clarify the compositional constraint that favor or impair the insertion of Low Complexity Regions. The reason behind the quint presence of the LC spread in *Laverania* and whether similar mechanisms rose in *Haemamoeba* remains obscure.

Table 1 Column 1 Correlation coefficients between protein length and LC residues Column 2 Average GC content Column 3 LC content

Organism	**P correlation coefficient**	**GC content**	**LC content (amino – acids count)**
P.falciparum	r = 0,70	0,23	130020
P.praefalciparum	r = 0,71	0,23	125491
P.reichnowi	r = 0,71	0,23	133878
P.alderi	r = 0,67	0,22	105572
P.gaboni	r = 0,70	0,22	102893
P.vivax	r = 0,35	0,48	28753
P.knowlesi	r = 0,26	0,40	19693
P.coatney	r = 0,32	0,44	15273
P.cynomolgi	r = 0,42	0,42	26101
P.inui	r = 0,26	0,44	16109
P.fragile	r = 0,22	0,44	19378
P.gonderi	r = 0,39	0,26	31857
P.yoeli	r = 0,33	0,24	29765
P.berghei	r = 0,38	0,24	28753
P.petteri	r = 0,21	0,27	11737
P.vinckei	r = 0,16	0,26	11324
P.chabaudi	r = 0,26	0,28	11265
P.gallinaceum	r = 0,46	0,21	49707
P.relictum	r = 0,37	0,21	39707
P.malariae	r = 0,45	0,30	31876
P.ovale	r = 0,43	0,36	43101
P.o.curtisi	r = 0,30	0,32	32005

3.2 Bonferroni Correction for LCP Protein Length Distributions

LCPs appear to be longer than nLCPs. In such a way, protein length appears to be a triggering feature for the insertion of LCRs. We wondered if any significant difference rose among LCP length distributions that justifies the whimsical LC content and correlations found in *Laverania subgenus* (i.e. we wondered if *Laverania* LCPs were significantly longer than the other LCPs to allow a major spread of LCRs). Therefore, we implemented a multi-comparison test (*Bonferroni Correction*) among LCP Protein length distributions depleted of their LC content. An interactive MATLAB plot showing this test is left in supplementary materials. We find any significant difference, as expected, between *P. falciparum* and the rest of *Laverania subgenus.* In turn, *P. falciparum* has significantly shorter LCPs than *P.gonderi* and *P.malariae* with *P.ovale* and *P.yoeli* showing shorter LCPs if compared with *P.falciparum*. *P.yoeli* flaunts the shortest LCPs among

all the considered organisms. Therefore, protein length alone, even appearing as a favorability feature for the insertion of LCRs, finally results as a necessary but not sufficient condition for a major spread of LCRs. Overall, no significant differences are found in *Laverania* that justify the whimsical behavior that characterizes this subgroup.

3.3 Amino Acid Composition of Low Complexity Strands

We left an *excel* file in Supplementary Materials. Each column represents an organism while each row represents an amino acid. The entire set of amino acids represent the amino acid distribution composing the LCRs of each organism. Each amino acid count is normalized over the total count of the residues composing LCRs of that organism (see Tab.1). Overall, we observe some differences among the various parasites represented by a few amino acid species. *Laverania* subgenus is characterized priorly by asparagine (N), as reported by other works with respect for *P. falciparum* (Pizzi & Frontali, 2001; Filisetti et al. 2013; Frugier et al.2009; Muralidharan & Golberg, 2013; Chaudhry et al. 2018). Asparagine represents the 54,3% of the total amino acid usage in each of these parasites on average. It is followed by Aspartic acid (D) (10,4% on average) and Glutamic Acid (E) (5,3% on average). As far as the other amino acids are concerned, we observe a minor expression (1-2%) with Serine (S) only, showing an averaged usage of nearly 4,7%. We find a different situation in *simian plasmodia* where the most expressed residues are D (~14,4%) and E (~22,2%). Asparagine results more rarely expressed (~8,4% on average). As before, S reach on average the 9,5% of the total count while the rest of the other amino acid species fluctuate between 1 and 3% on average. In *Haemamaeba*, as expected given their closer phylogenetic relationship with *P.falciparum* (Waters et al. 1991; Waters et al. 1993 a-b; McCutchan et al. 1996;Escalante & Ayala,1994), we observe a preference for N (~ 38,6% on average) followed by D and E (~9,2% and ~20,2% respectively). Conversely, S represents nearly the 5,2% on average with the other amino acids varying from 1 to 2-3% on average. *Vinckeia* despite their phylogenetic relationship observed with *P. falciparum* (Yadav & Swati, 2012) dispose a rather distant situation. D and E are the most used residues among LCRs (12,3 % and 20 % on average respectively) while asparagine disposes some fluctuations with only *P.yoeli* and *P.berghei* with over the 30% of N residues in their LCRs (28,6 % and 35,5% respectively). *P.vinckei*, *P.chabaudi* and *P.petteri* express between the 16% and 13% of asparagine residues (~16,5%, 13,37% and 16,6% respectively). Serine reaches the 6,6% on average. The LC amino acid composition of OHIP is sprinkled with D, E, N and S representing the 11,23,10 and 12 % respectively. For further details we refer to the excel file. A second excel file with the count of each amino acid species belonging to the LCRs of each organism is left in Supplementary.

3.4 ENC

The effective number of codons (ENC) is a widely used index for characterizing codon usage bias because it does not require a set of reference genes as does, *e.g.*, codon adaptation index (CAI) (Sharp & Li, 1987). We relied on its improved implementation proposed by Sun and colleagues (2013). As originally formulated (Wright, 1990), the theoretical curve represents those ENC values that indicate those proteins where the synonymous codons are chosen with no-codon preference. CUB for these proteins must be attributed entirely to mutational bias. On the other hand, the farther the distance from the WTC the higher the meddling of selective pressure in shaping codon usage bias (Novembre, 2002). Fig 2,3,4,5 and 6 show the ENC plots for *Laverania*, *simian plasmodia*, *Vinckeia*, *Haemamaeba* and OHIP, respectively. In red we represented LCPs, in green nLCPs. At first glance, LCPs show to be closer to the WTC. This characteristic appears in all the subgroups despite the extremely wide variance of ENC and GC_3 values. *Laverania*, *Vinckeia* and *Haemamaeba* show to be extremely AT rich with their ENC distributions placed on the left side of the plots together with OHIP despite their GC_3 content appears less extremized. *Simian plasmodia* place their distribution toward the center of the ENC plane. We implemented a set of correlation analysis (*Pearson*) between ENC and GC_3 values for the AT biased plasmodia. A correlation between ENC and GC_3 in *P. falciparum* was first spotted in other works (Gajbhiye et al. 2017). Results are collected in Tab.2. The first column collects the statistics for nLCPs while the second the statistics for LCPs. Overall, these species show strong linear correlations for both the LCPs and nLCPs. Synonymous codons appear to increase as the GC_3 increases as well. The correlation coefficients of nLCPs and LCPs flaunt similar values and no significant difference are found between the two correlation distributions (*Mann-Whitney*, $p>0,5$). All the correlations are significant ($p<<0,01$). Tab. 3 collects the statistics for *simian plasmodia*. We wondered if

different kind of relations were between ENC and GC_3 content. We implemented fits with polynomial of second order. All the correlations are significant (*p<<0,01*). ENC values in *simian plasmodia* appear to reach their maximum near GC_3 = 0,5 decreasing for greater values. In particular *P.vivax* see its most GC rich genes falling at greater distance from the WTC that implies a major strength of purifying selection. Differently from the rest of its clade, *P.gonderi* show a rather biased AT content with strong linear correlation between ENC and GC_3 value where the related values are reported in Tab.2.

Table 2 Correlation coefficients between GC_3 content and ENC in nLCPs and LCPs for *Laverania, Vinckeia, Haemamaeba and OHIP*

Organism	**nLCP**	**LCP**
P.falciparum	r = 0,56	r = 0,65
P.prafalciparum	r = 0,57	r = 0,62
P.reichnowi	r = 0,62	r = 0,62
P.alderi	r = 0,61	r = 0,60
P.gaboni	r = 0,59	r = 0,53
P.yoeli	r = 0,54	r = 0,52
P.berghei	r = 0,57	r = 0,62
P.petteri	r = 0,54	r = 0,47
P.vinckei	r = 0,57	r = 0,54
P.chabaudi	r = 0,60	r = 0,51
P.gallinaceum	r = 0,42	r = 0,43
P.relictum	r = 0,41	r = 0,41
P.malariae	r = 0,68	r = 0,72
P.ovale	r = 0,65	r = 0,59
P.o.curtisi	r = 0,78	r = 0,82
P.gonderi	r = 0,67	r = 0,71

Despite its utility ENC plot alone is poorly informative and to test whether LCPs are expressed by a greater number of codons than nLCPs we implemented a pairwise comparison between LCPs and nLCPs as we did for protein length distributions. Results are collected in Fig.7 (*welch t-test*).

Table 3 **Goodness of fit between** GC_3 and ENC in simian plasmodia

Organism	**nLCP**	**LCP**
P.vivax	r-square = 0,67	r-square = 0,50
P.knowlesi	r-square = 0,23	r-square = 0,47
P.inui	r-square = 0,20	r-square = 0,35
P.cynomolgi	r-square = 0,56	r-square = 0,34
P.coatney	r-square = 0,32	r-square = 0,42
P.fragile	r-square = 0,53	r-square = 0,59

In most of the cases, LCPs tend to be expressed with a higher number of codons that is synonymous of a more lenient purifying selection. LCPs in *Laverania* are expressed with greater ENC values (*p<<0,01*) than nLCPs and similar results are found in *simian plasmodia* (*p<<0,01*) as well as it is for *Haemameaba P.ovale, P.malariae* and *P.o.curtisi* (*p<<0,01*). Controversial, is the situation in *Vinckeia* where *P.chabaudi, P.yoeli* and *P.berghei* show any significant distinction between LCPs and nLCPs. Eventually, Fig 7 allows to observe the evolutionary diversification of these species of parasites.

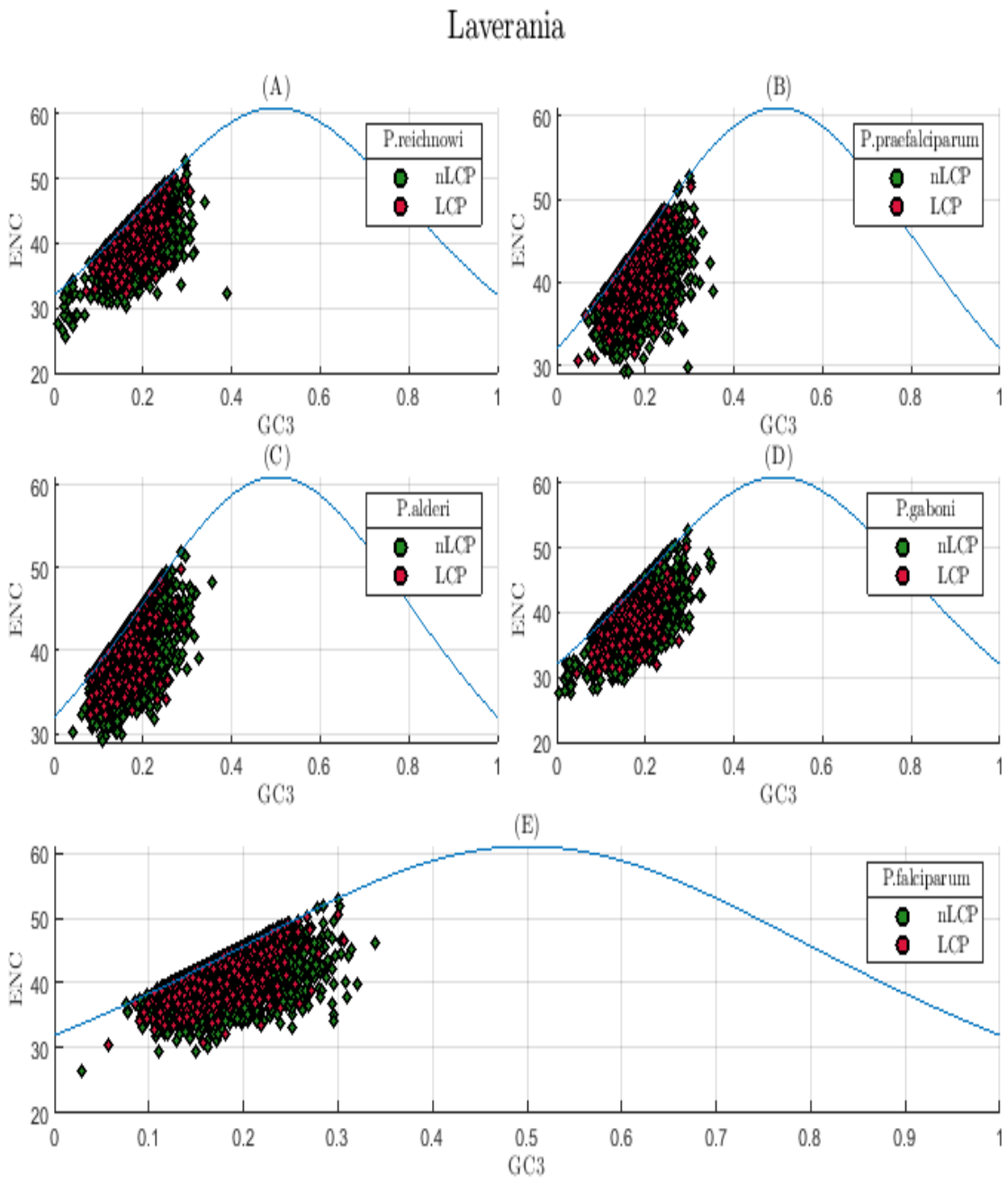


Fig 1 *ENC plot for Laverania sub genus (A) P.reichnowi (B) P.praefalciparum (C) P.alderi (D) P.gaboni (E) P.falciparum*

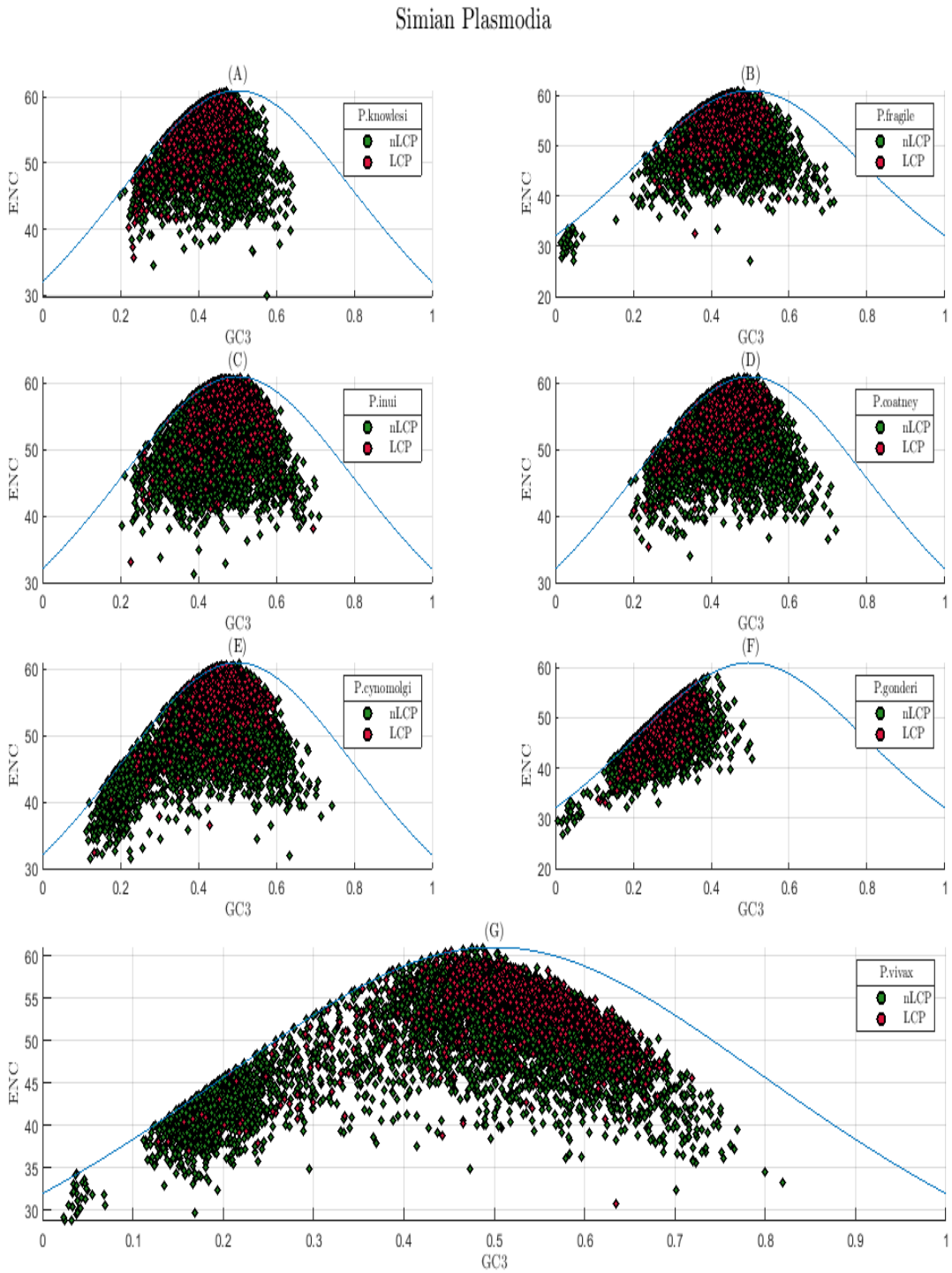


Fig 2 *ENC plot for simian plasmodia (A) P.knowlesi (B) P.fragile (C) P.inui (D) P.coatney (E) P.cynomolgi (F) P.gonderi (G) P.vivax*

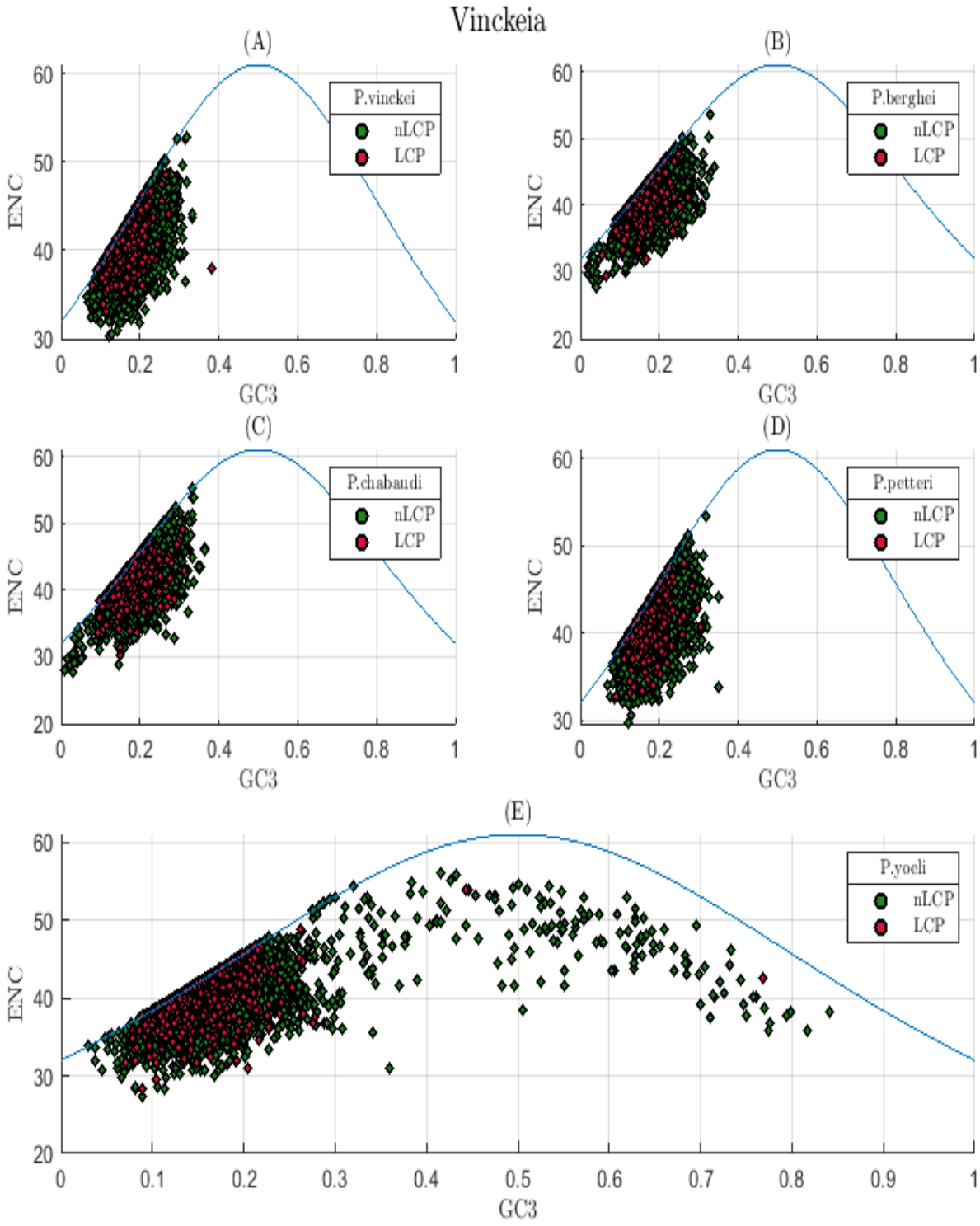


Fig 3 *ENC plot for Vinckeia subgenus (A) P.vinckei (B) P.berghei (C) P.chabaudi (D) P.petteri (E)P.yoeli*

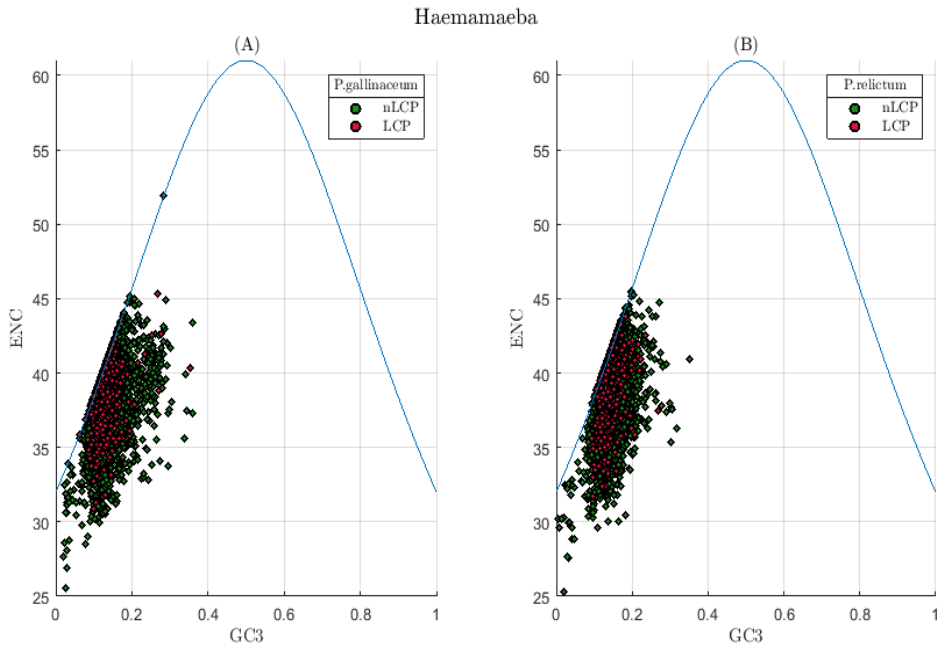


***Fig** 4 ENC plot for Haemamaeba subgenus (A) P.gallinaceum (B) P.relictum*

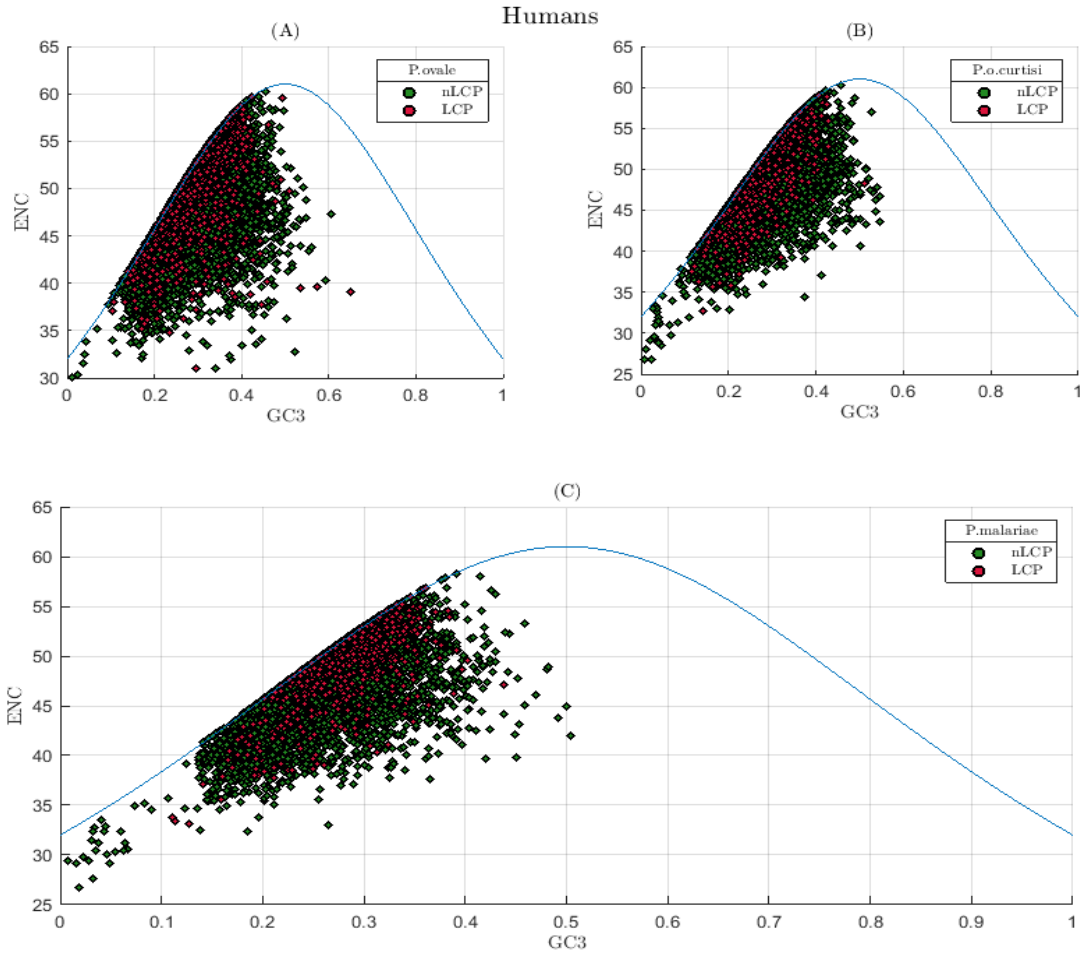


***Fig 6** ENC plot for (A) P.ovale (B) P.o.curtisi (C) P.malariae*

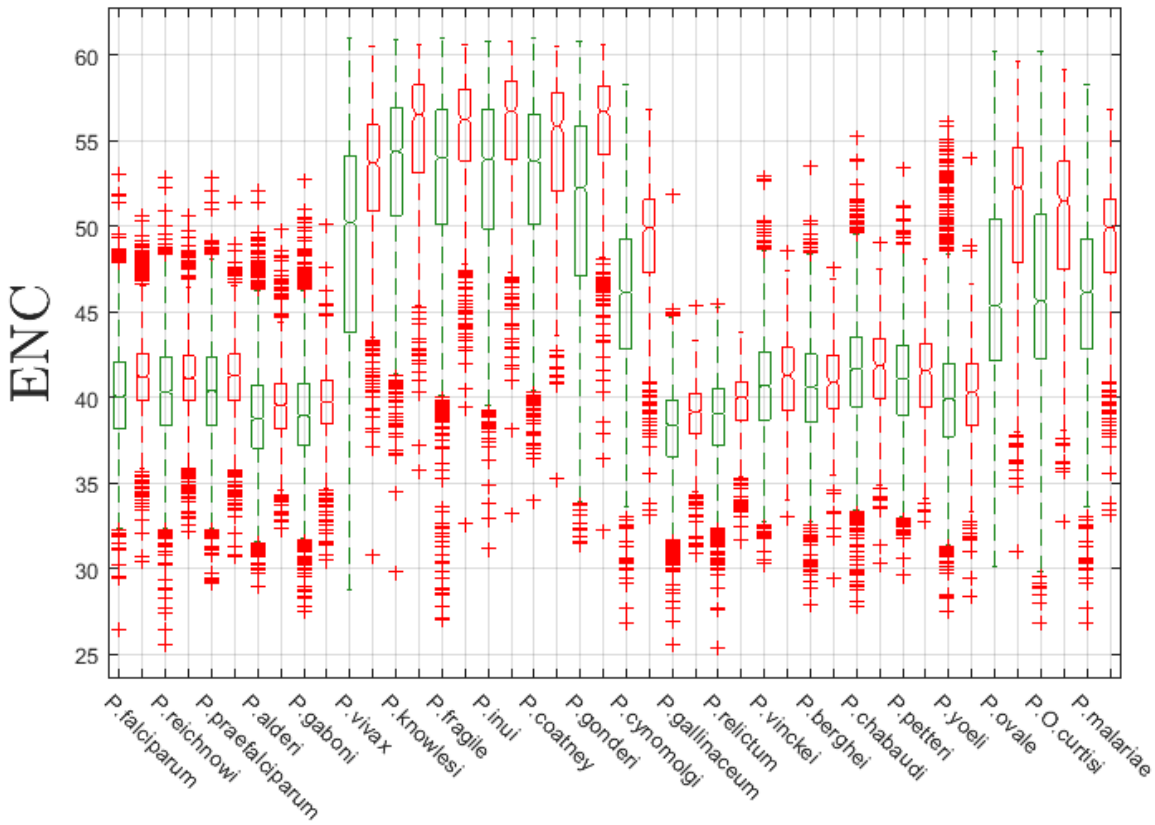


Fig 7 *ENC distributions for all the considered plasmodia. For each pair of box plots in red LCPs' ENC distribution while in green we represented nLCPs'. All the pairwise comparison are significant (p<<0,00001). However, some of the vinckeia do not show these distinctions (p>0,5).*

3.5 SPI

ENC plot provided important details about CUB showing how, on average, LCPs tend to be expressed by a greater number of codons. Mutational bias appears to act as the priority body in shaping the CUB in LCPs. However, ENC values have some problems since the same ENC value can correspond to a very different action of the selective pressure given a different GC content. SPI helps to solve this riddle measuring the distance of each gene from the curve (see Method). Thus, a SPI value of 0 will correspond to a situation where CUB is decided entirely by mutational pressure (situation of no bias). When SPI is instead equal to 1 selective pressure acts as the primary body in shaping CUB (situation of extreme bias). We implemented, as we did for protein length and ENC distributions, a pairwise comparison showed in Fig.8. Results are coherent and illustrate in a better fashion the role of both mutational and selective pressure. nLCPs tend to be farther from the WTC while LCPs tend to depart the less. Therefore, the action of selective pressure results to be preeminent on nLCPs. Instead, even if selective pressure operates in a not negligible way, CUB in LCPs tends to be shaped by Mutational Bias with priority. Each pairwise comparison is significant (*p<<0,01* on average).

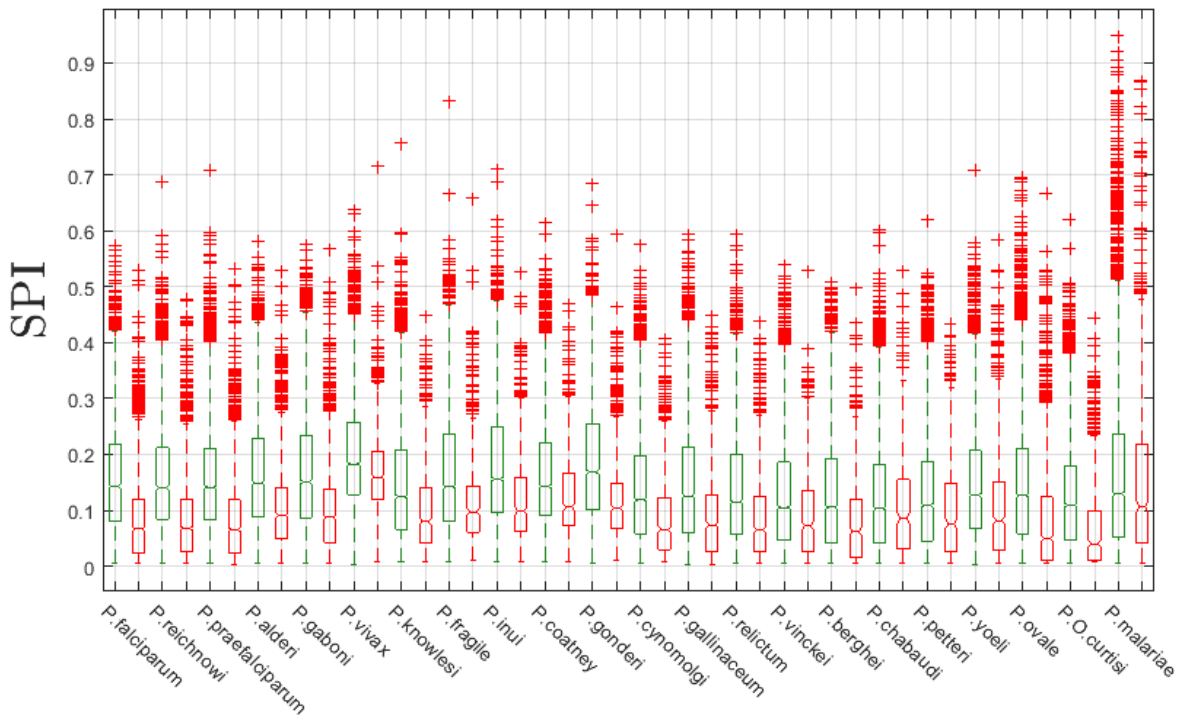


Fig 8 *SPI distributions. In red LCPs in green nLCPs. LCPs are nearer the expected curve since their SPI values tend to zero while nLCPs tend to depart more from the expected curve. All the pairwise comparisons are significant (p<<0,00001)*

3.6 Bonferroni correction for LCP SPI distributions

Mutational bias appears to be one of the requisites to allow LCRs to squeeze into interdomains. In this sense, the pauperism of mutational constraints appears to favour these insertions. Furthermore, the mutational bias appears to be an evolutionary signal indicating a LCPs' tendency to acquire mutations without impairing their function. We wondered if any significant distinction rose in SPI distributions such as to justify the LC overabundance that characterizes *Laverania subgenus.* Overall, we find no significant distinctions such as to clarify the major spread of LC in L*averania.* This result shows that mutational bias, despite being a triggering factor, cannot explain alone why this characteristic is accentuated in one organism rather than in another. Considering *P. falciparum, P.reichnowi* and *P.praefalciparum* they flaunt smaller SPI values with respect to *P.alderi and P.gaboni*. Moreover, differently from these latter, the former trio disposes a set of SPI values that can be compared with *simian plasmodia,* made exception for *P.vivax* that shows the greatest SPI values among all the analysed organisms. Always considering *P.falciparum, P.reichnowi* and *P.praefalciparum,* their respective SPI distributions do not differ with significance from the most of *vinckeia* (made exception for *P.chabaudi, P.yoeli* and *P.petteri* that flaunt higher SPI values that in turn are comparable with *P.alderi* and *P.gaboni*) and from *Haemamaeba,* with *P.ovale* and *P.o.curtisi*. On the other hand, *P. malariae* exhibits a SPI distribution for its LCPs that places itself before *P.vivax* with significance. Thus, it appears that mutational bias alone does not explain why an organism can be characterized by a major or poorer content in LC. In general, it appears that a lack of mutational constraints allows the insertion of LCRs. An interactive MATLAB plot with the Bonferroni Correction is left in the Supplementary Materials.

Discussion

It is useful to establish a general informative background in order to clarify our discussion. Starting from the selection-mutation drift theory (Bulmer, 1991) it is generally accepted that strong biases in CUB would occur in highly expressed genes, while in weaker ones the persistence of non-optimal codons would be allowed together with the former category due to selection for speed of translation or translational efficiency. Therefore, the selection is likely to be stronger on codons in highly expressed genes since they are used more often (Bulmer 1987). Moreover, selection has an important role in determining protein length where shorter proteins would improve the fitness minimizing the cost of translation by the elimination of extra sequences that do not introduce any significative improvement in fitness or minimizing the cost in term of chaperones utilized to fold the protein (Lipman et al.2002; Hartl et al. 2002). Of great interest, deletions appear to be more frequent than insertion (Lipman et al. 2002). This implies that in the economy of selection shorter proteins appear to have undergone more evolutionary steps, being passed through several actions of the natural selection that finally chose those patterns of codons that allowed the organisms to better survival, reproduce finally optimizing their internal biosynthetic ecosystem, being filtered of their extra (luckily non-coding) DNA (Cavalier,1982). On the other hand, comparisons of set of orthologous proteins enlightened that "longer" proteins appear to be emerged in more recent evolutionary times (Wang et al. 2005). Furthermore, newly evolved proteins show to have a greater content in LCRs (Toll-Riera et al. 2011) where primordial proteins are hypothesised to be highly repetitive oligomers that specified periodical polypeptides (Ohno & Epplen,1983). Coherently, these repetitive sequences are recognized to be under low selective pressure providing thus a ground for experimentation for natural selection (Pancsa & Tompa, 2016), gradually developing new functions and optionally increasing the complexity of the sequence leading to new protein folds. Conversely, LCRs are found in performing some important biological functions (Shen et al.2004; Salichs et al.2009) making logical to grant a certain weight in maintaining the fitness of the organisms. This implies a major role of the selective pressure in preserving these regions. Otherwise they can also vanish (Pancsa & Tompa, 2016). Our results can be summarized as follows: (i) the first pairwise comparison showed that LCPs are longer than nLCPs. (ii) The correlations between protein length and LC content showed a special affinity for LCRs in *Laverania subgenus* where such an affinity is not observed in all the other subgroups. We depleted LCPs (iii) from their LC content. The LCPs resulted to be longer than nLCPs despite this depletion. These two results confirmed the LCPs to be longer than nLCPs *de facto* and that protein length is one of the causative features responsible of the presence of Low Complexity stretches. (iv)As already stressed by other (Pizzi & Frontali, 2001), We observed that the LC insertion unlikely depends from GC. In fact, *Vinckeia subgenus* despite being A-rich cluster is not as LC loaded as *Laverania* is. Overall, it appears that even if the a genome phenotype property as the A-richness of the coding strands condition the classical phenotype in the choice of the amino acids, GC content does not appear as the primary body in influencing the insertion of LCRs. Moreover, we wondered if the protein length of LCPs in *Laverania subgenus* could have possibly influenced the insertion of LCRs searching for a difference from the other LCPs. The *Bonferroni correction* (v) reveals any significant difference among these distributions with some minor LC containing organisms with longer LCPs than *Laverania*. Protein Length, despite appearing as a triggering factor for the insertion of Low Complexity strands, does not explain completely the motivation behind the whimsical characteristics of *Laverania plasmodia* to further generalize the phenomenon of the LC insertion. The *Mann-Whitney test* (vi) enlightened a significant difference in LC content among the subgenus of *Laverania*, *Vinckeia* and *Simian Plasmodia*. *Haemamoeba* and OHIP do not show a significant difference from *Laverania*, even if the specimen pauperism decreases our confidence underlining the necessity to have more organisms to analyse. (vii) The observation of the amino acid content enlightened a sort of preference for asparagine in *Laverania*. This subgenus appears unique since the other species do not show similar preference in the addition of a singular amino acid species. On the other hand, D, E and S appear to be three commonly LC expressed amino acids in all the species here analysed. (viii) The analysis of the ENC distributions casted some light about the role of selection upon plasmodium proteins. At first glance, the mere observation of the ENC plots showed that LCPs appear to depart the less from the WTC in all the analysed plasmodia. We observed two scenarios (*pairwise comparison*): nLCPs manifest a more stringent CUB while LCPs flaunt a more flexible codon repertoire being expressed, on average, by a ENC values. This suggested how the purifying pressure acts primarily on nLCPs while the mutational

pressure, instead, shapes the CUB of LCPs with priority. The same boxplot enlightens the evolutionary diversification of these organisms. In this sense, a parallelism results to be useful. HIV1 is one of the most extreme AT-rich viruses (Hemert et. Al 2001) while HTLVI is extreme in the GC content. By these two properties the genomes of HIV and HTLV 1/2 mimic an AG-rich and GC-rich segment of the host genome, respectively (Kypr et al. 1989). Presuming they both arose from a common ancestor, the same pressure would luckily have caused prototypic malarial parasite lineages to diverge, maybe deriving from a common AT-rich ancestor (Nikbakht et al. 2014), further adapting to new host species. This make us infer that as adaptations to their new host species grew more complete, then the pressure to differ in base composition from other species would have decreased, since there was no longer a possibility of intracellular genome blending. Following, (ix) with no significative differences between nLCPs and LCPs in AT biased plasmodia comprising also *P.gonderi,* we find linear correlations between ENC values and GC_3. The usage of synonymous codons goes toward a situation of no-codon preference as the GC_3 increases, appearing in a similar fashion in both LCPs and nLCPs. *Simian plasmodia* showed different rules where the ENC vs GC_3 distributions follow quadratic fits. Overall, for *simian plasmodia,* we observed that the CUB is more demanding at the extremes of the ENC planes while more lenient toward the centre of the graph. (ix) Given some problems inherent to the ENC we proposed the SPI. In fact, some of the plasmodia of *Vinckeia subgenus* failed in the ENC pairwise comparison, probably due to the redundancy of the ENC values (see methods). SPI solves these problems enlightening how, with statistical significance, nLCPs undergo a more stringent action of the purifying selection (they are closer to an extreme bias situation), depicting instead a second symmetric scenario where we find LCPs to be closer to the model of no-codon preference (no bias). Finally, mutational bias appears one of the triggering factors that allows the insertion of LCRs in *plasmodia.* Once more, we wondered if the balance between the two evolutionary forces shifted in a whimsical fashion in *Laverania* such as to justify their *"taste"* for Low Complexity. The following *Bonferroni correction* (x) revealed any significant insights observing that the most plasmodia have similar SPI distributions. Comprehensively, we can draw a synthesis: our results bring us to attribute LCPs and nLCPs to different evolutive scenarios. nLCPs appear to be under a more demanding action of the purifying pressure suggesting a difficulty to accept mutations that could impair their primary function. Furthermore, the extreme difference in length with respect to LCPs make us infer that these proteins underwent more actions, maybe passing from a speciation event to another, of the natural selection that rewarded more compact synonymous codon patterns, presumably to enhance their translation rates removing extra DNA sequences that did not introduce any significant benefit. We thus propose these proteins as the main protein asset for the parasite survival. Symmetrically, the characteristics of LCPs underline a lack of structural definition implying a situation where selection could redefine the nucleic acid asset with more freedom. These mutations would not impair necessarily the fitness of these organisms. In this sense, and given the quint protein length, LCPs appear as recently acquired. To answer at our first question, in general, the replication slippage would be promoted by a lack of evolutionary constraints (Mutational Bias acts with priority on LCPs) luckily providing a source of mutability and adaptation for these parasites where the selection for a few amino acids species would provide repetitive strands of one amino acid type rather than another.

Conclusion

We profiled some important characteristics of plasmodia proteins depicting two scenarios seeing how LCPs could be an evolutionary and adaptive asset for *plasmodia* given the more relaxed action of selective pressure and whimsical characteristics that permit to LC domains to spread all over the aminoacidic chains. On the other hand, nLCPs appear to be the engine by which these organisms establish their core biological behaviour given the more stringent action of selective pressure that makes nLCPs less prone to accept mutations. Overall, our results provide new insight towards general laws of protein evolution identifying the LCPs proteins as a reservoir for evolutionary innovations and nLCPs as the core asset for parasites' fitness. Quantifying selective and mutational forces acting on plasmodia genes could be useful for future studies concerning protein *de novo* design, synthetic biology, biotechnology applications, and the identification of proteins that are relevant to strike down malaria worldwide.

Authors Contributions: A.C. proposed the SPI; A.C. proposed the analysis; A.G. proposed the depletion of LC in the length distribution pairwise comparison; A.C. implemented the code; A.C. analyzed the results; A.C. wrote the text. Each author, discussed and accepted the text manuscript

List of Abbreviations: LCPs, Low Complexity Containing Proteins; RS, Replication Slippage; nLCPs, not Low Complexity Containing Proteins; LCRs, Low Complexity Regions; LCSs, Low Complexity Sequences; LC, Low Complexity; OHIP, Other Human Infecting Plasmodia

Conflict of Interest: Authors declare no conflict of interest

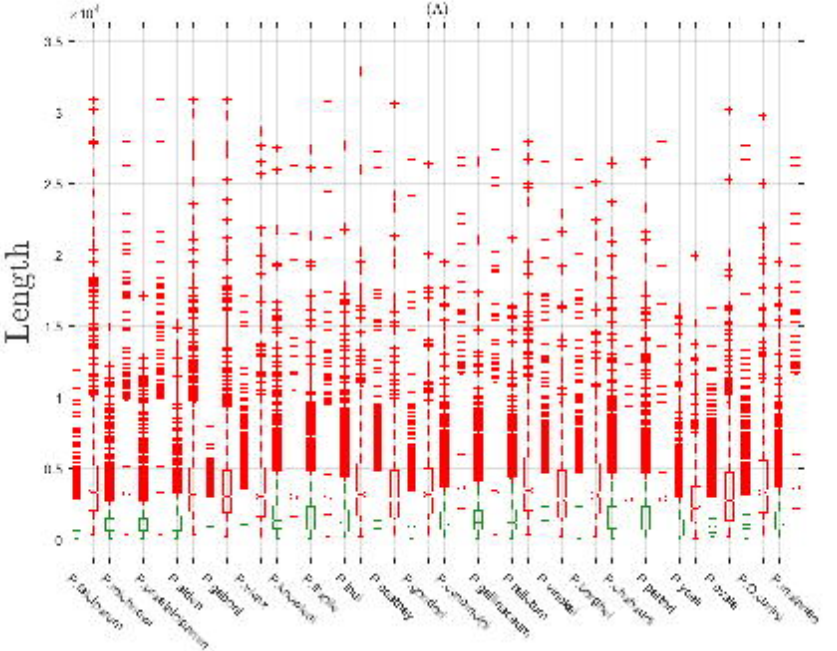
Funding: This research received any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

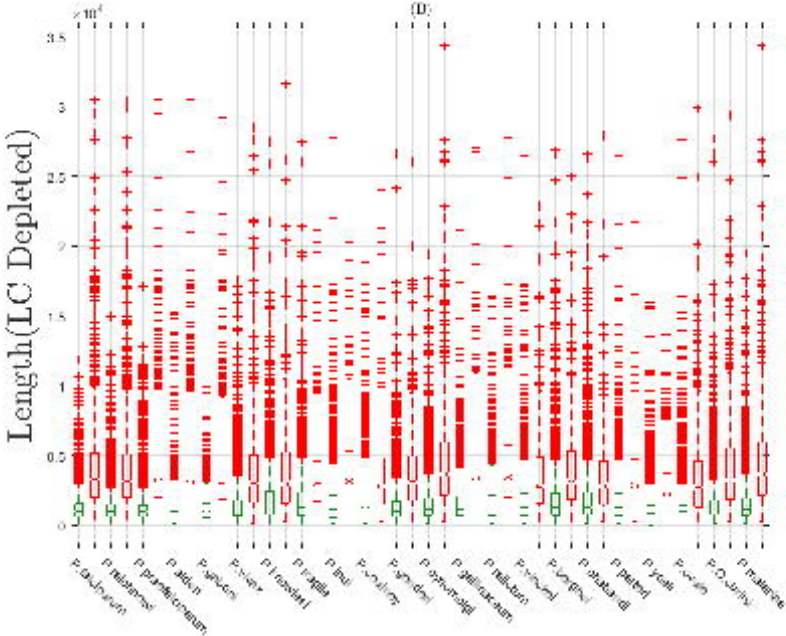
References

- Arisue, N., Hashimoto, T., Kawai, S. *et al*. Apicoplast phylogeny reveals the position of *Plasmodium vivax* basal to the Asian primate malaria parasite clade. *Sci Rep* **9,** 7274 (2019). https://doi.org/10.1038/s41598-019-43831-1
- Batistuzzi FU, Schneider KA, Spencer MK, Fisher D, Chaudhry S, Escalante AA. Profiles of low complexity regions in Apicomplexa. BMC Evol Biol. 2016;16:47. Published 2016 Feb 29. doi:10.1186/s12862-016-0625-0
- Bernardi G. and Bernardi G. (1986a), Compositional Constraints and Genome Evolution, J. Mol. Evol. 24: 1-11
- Brown LY, Brown SA. Alanine tracts: the expanding story of human illness and trinucleotide repeats, Trends Genet., 2004, vol. 20 (pg. 51-58)
- Bulmer M. Coevolution of codon usage and transfer RNA abundance. Nature. 1987 Feb 19;325(6106):728–730
- Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129(3):897-907.
- Carter, R. and 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
- Cavalier-Smith T. (ed) The evolution of Genome Size, Wiley New-York,1985
- Chaudhry, S.R., Lwin, N., Phelan, D. *et al*. Comparative analysis of low complexity regions in Plasmodia. *Sci Rep* **8,** 335 (2018). https://doi.org/10.1038/s41598-017-18695-y
- Collins WE, Jeffery GM. Plasmodium malariae: parasite and disease. *Clin Microbiol Rev*. 2007;20(4):579-592. doi:10.1128/CMR.00027-07
- Collins WE, Jeffery GM. Plasmodium ovale: parasite and disease. *Clin Microbiol Rev*. 2005;18(3):570-581. doi:10.1128/CMR.18.3.570-581.2005
- Corradetti A.; Garnham P.C.C.; Laird M. (1963). "New classification of the avian malaria parasites". Parassitologia. 5: 1–4.
- Ellegren H. Microsatellites: simple sequences with complex evolution, Nat Rev Genet., 2004, vol. 5 (pg. 435-445)
- Escalante AA, Ayala FJ. Phylogeny of the malarial genus Plasmodium derived from rRNA gene sequences. Proc Natl Acad Sci U S A. 1994;91(24):11373-11377. doi:10.1073/pnas.91.24.11373
- Escalante AA, Pacheco MA. Malaria Molecular Epidemiology: An Evolutionary Genetics Perspective. MicrobiolSpectr. 2019;7(4): 10.1128/microbiolspec.AME-0010-2019. doi: 10.1128/microbiolspec.AME-0010-201
- Escalante, Ananias & Freeland, Denise & Collins, William & Lal, Altaf. (1998). The Evolution of Primate Malaria Parasites Based on the Gene Encoding Cytochrome b from the Linear Mitochondrial Genome. Proceedings of the National Academy of Sciences of the United States of America. 95. 8124-9. 10.1073/pnas.95.14.8124.
- Faust, Christina & Dobson, Andrew. (2015). Primate malarias: Diversity, distribution and insights for zoonotic Plasmodium. One Health. 1. 10.1016/j.onehlt.2015.10.001.
- Filisetti, D., Th´eobald-Dietrich, A., Mahmoudi, N., Rudinger-Thirion, J., Candolfi, E., Frugier, M. (2013). Aminoacylation of Plasmodium falciparum tRNA Asn and Insights in the Synthesis of Asparagine Repeats. Journal of Biological Chemistry, 288(51), 36361–36371. doi:10.1074/jbc.m113 .522896

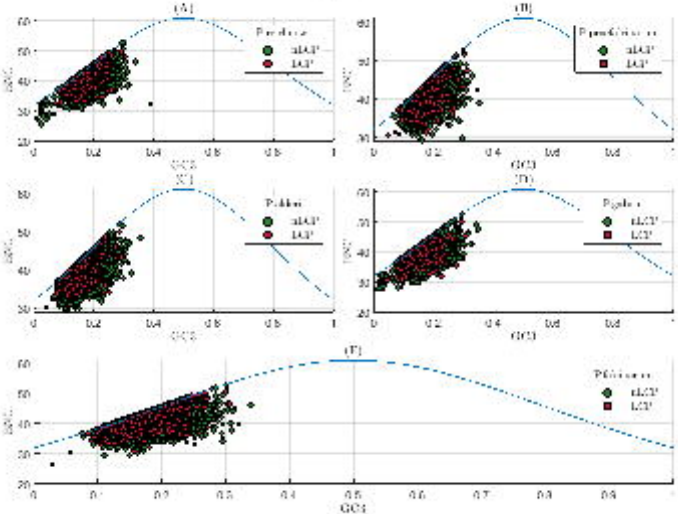
- Forcelloni, S., Giansanti, A. Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. J Mol Evol (2019) doi:10.1007/s00239-019-09921-4 Fowler,
- Forsdyke D., Evolutionary Bioinformatics 2016, Springer Ed.
- Frugier, M., Bour, T., Ayach, M., Santos, M. A. S., Rudinger-Thirion, J., Th´eobald-Dietrich, A., Pizzi, E. (2009). Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. FEBS Letters, 584(2), 448–454. doi:10.1016/j.febslet.2009.11.004
- Gajbhiye, Shivani & Patra, P.K. & Yadav, Dr. Manoj. (2017). New insights into the factors affecting synonymous codon usage in human infecting Plasmodium species. Acta Tropica. 176. 10.1016/j.actatropica.2017.07.025.
- Galinski M.R., Barnwell J., 2012, Nonhuman primate models for human malaria research. In:Abee, C.R. (Ed.), Nonhuman Primates In Biomedical Research: Diseases, Academic Press Elsevier, pp 299-323
- Galland G.G., Role of the Squirrel Monkey in Parasitic Disease Research, *ILAR Journal*, Volume 41, Issue 1, 2000, Pages 37–43, https://doi.org/10.1093/ilar.41.1.37
- Garnham PC. The Subgenera of Plasmodium In Mammals. *Ann Soc Belges Med Trop Parasitol Mycol*. 1964;44:267-271.
- Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles, Nat Rev Genet., 2005, vol. 6 (pg. 743-755)
- Green H, Wang N. Codon reiteration and the evolution of proteins. Proc Natl Acad Sci U S A. 1994;91(10):4298–4302. doi:10.1073/pnas.91.10.4298
- Mosbach V, Poggi L, Richard GF. Trinucleotide repeat instability during double-strand break repair: from mechanisms to gene therapy. Curr Genet. 2019;65(1):17-28. doi:10.1007/s00294-018-0865-1.
- Guy-Franck R., Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008;72(4):686–727. doi:10.1128/MMBR.00011-08 15
- Halfmann R, Alberti S, Krishnan R, Lyle N, O'Donnell CW, et al. (2011) Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. Mol Cell 43: 72–84.
- Hartl FU, Hayer-Hartl M. Molecular chaperones in the cytosol: from nascent chain to folded protein. Science. 2002;295:1852–1858. doi: 10.1126/science.1068408.
- Howes RE, Battle KE, Mendis KN, et al. Global Epidemiology of Plasmodium vivax. Am J Trop Med Hyg. 2016;95(6 Suppl):15-34. doi:10.4269/ajtmh.16-0141
- John Federhen, Scott. (1993). Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. Computers Chemistry. 17. 149- 163. 10.1016/0097-8485(93)85006-X. Xiao H, Jeang KT. Glutamine-rich domains activate transcription in yeast Saccharomyces cerevisiae, J Biol Chem. , 1998, vol. 273 (pg. 22873-22876)
- Kashi, Y., King, D. (2006). Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics, 22(5), 253–259. doi:10.1016/ j.tig.2006.03.005
- Kristan M, Thorburn SG, Hafalla JC, Sutherland CJ, Oguike MC. Mosquito and human hepatocyte infections with Plasmodium ovale curtisi and Plasmodium ovale wallikeri. *Trans R Soc Trop Med Hyg*. 2019;113(10):617-622. doi:10.1093/trstmh/trz048
- Kumari, B., Kumar, R., Kumar, M. (2015). Low complexity and disordered regions of proteins have different structural and amino acid preferences. Molecular BioSystems, 11(2), 585–594. doi:10.1039/c4mb00425f
- Kypr J, Mrázek J, Reich J. Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Biochim Biophys Acta. 1989;1009(3):280-282. doi:10.1016/0167-4781(89)90114-0
- Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. The relationship of protein conservation and sequence length. *BMC Evol Biol*. 2002; 2:20. doi: 10.1186/1471-2148-2-20
- Liu W, Sundararaman SA, Loy DE, et al. Multigenomic Delineation of Plasmodium Species of the Laverania Subgenus Infecting Wild-Living Chimpanzees and Gorillas. *Genome Biol Evol*. 2016;8(6):1929-1939. Published 2016 Jul 2. doi:10.1093/gbe/evw128
- McCutchan TF, Kissinger JC, Touray MG, Rogers MJ, Li J, Sullivan M, Braga EM, Krettli AU, Miller LH. Comparison of circumsporozoite proteins from avian and mammalian malarias: biological and phylogenetic implications. Proc Natl Acad Sci USA. 1996;93:11889–11894.
- Mitsui H., Arisue N, Sakihama N, Inagaki Y, Horii T, Hasegawa M, Tanabe K, Hashimoto T., Phylogeny of Asian primate malaria parasites inferred from apicoplast genome-encoded genes with special emphasis on the positions of Plasmodium vivax and P. fragile, Gene. 2010 Jan 15;450(1-2):32-8. doi: 10.1016/j.gene.2009.10.001
- Muralidharan V, Goldberg DE. Asparagine repeats in Plasmodium falciparum proteins: good for nothing? PLoS Pathog. 2013;9(8):e1003488. doi:10 .1371/journal.ppat.1003488
- Muralidharan V, Oksman A, Pal P, Lindquist S, Goldberg DE. Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers. Nat Commun. 2012; 3: 1310. doi: 10.1038/ncomms2306

- Nikbakht, H., Xia, X., & Hickey, D. A. (2014). *The evolution of genomic GC content undergoes a rapid reversal within the genus Plasmodium. Genome, 57(9), 507–511.* doi:10.1139/gen-2014-0158
- Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394
- Ohno, S., Epplen, J. T. (1983). The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proceedings of the National Academy of Sciences, 80(11), 3391–3395. doi:10.1073/pnas.80.11.3391
- Otto TD, Gilabert A, Crellen T, et al. Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nat Microbiol*. 2018;3(6):687–697. doi:10.1038/s41564-018-0162-2
- Pancsa R., Tompa P., Coding Regions of Intrinsic Disorder Accommodate Parallel Functions, Published:September 16, 2016.DOI:https://doi.org/10.1016/j.tibs.2016.08.009
- Pizzi E., Frontali C. (2001). *Low-Complexity Regions in Plasmodium falciparum Proteins. Genome Research, 11(2), 218–229.* doi:10.1101/gr.gr-1522r
- Prugnolle F, Durand P, Neel C, et al. African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. *Proc Natl Acad Sci U S A*. 2010;107(4):1458-1463. doi:10.1073/pnas.0914440107
- Rich SM, Xu G. Resolving the phylogeny of malaria parasites. Proc Natl Acad Sci U S A. 2011;108(32):1297312974. doi:10.1073/pnas.1110141108
- Salichs E, Ledda A, Mularoni L, Alba MM, de la Luna S. 2009. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. PLoS Genet. 5:e1000397
- Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281D1295. doi: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. Mol Cell. 13:367–376.
- Sibley D.L., Howlett B.J., Heitman J., Evolution of Virulence in Eukaryotic Microbes, SBN: 978-1-118-30813-4, June 2012, Wiley-Blackwell
- Singh B, Daneshvar C. Human infections and detection of Plasmodium knowlesi. Clin Microbiol Rev. 2013;26(2):165-184. doi:10.1128/CMR.00079-12
- Sun X, Yang Q, Xia X. An improved implementation of effective number of codons (nc). *Mol Biol Evol*. 2013;30(1):191-196. doi:10.1093/molbev/mss201
- Toll-Riera, M., Rado-Trilla, N., Martys, F., Alba, M. M. (2011). Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. Molecular Biology and Evolution, 29(3), 883–886. doi:10.1093/ molbev/msr263
- Valkiunas G., Avian Malaria Parasites and Other Haemosporidia, CRC Press, 2000 NW Corporate Blvd., Boca Raton, Florida, 33431 USA. 2005. 932 pp. ISBN 0- 415-30097-5.
- van Hemert FJ, van der Kuyl AC, Berkhout B. The A-nucleotide preference of HIV-1 in the context of its structured RNA genome. RNA Biol. 2013;10(2):211-215. doi:10.4161/rna.22896
- Videvall E. Plasmodium parasites of birds have the most AT-rich genes of eukaryotes. *Microb Genom*. 2018;4(2):e000150. doi:10.1099/mgen.0.000150
- Vogel G (November 2013). "The forgotten malaria". Science. 342(6159): 684–7. doi:10.1126/science.342.6159.684
- Wang D, Hsieh M, Li WH. A general tendency for conservation of protein length across eukaryotic kingdoms. *Mol Biol Evol*. 2005;22(1):142-147. doi:10.1093/molbev/msh263
- Waters AP, Higgins DG, McCutchan TF. Evolutionary relatedness of some primate models of *Plasmodium*. Mol Biol Evol. 1993b; 10: 914–923.
- Waters AP, Higgins DG, McCutchan TF. The Phylogeny of malaria: a useful study. Parasitol Today. 1993a;9:246–250
- Waters, A. P., D. G. Higgins, and T. F. Mccutchan. 1991. Plasmodium falciparum appears to have arisen as a result of lateral transfer between avian and human hosts. Proceedings of the National Academy of Sciences USA 88: 3140-3144
- Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*. 1996; 266:554-571. doi:10.1016/s0076-6879(96)66035-2
- Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990;87(1):23-29. doi:10.1016/0378-1119(90)90491-9
- Yadav MK, Swati D. Comparative genome analysis of six malarial parasites using codon usage biasbased tools. *Bioinformation*.2012;8(24):1230-1239. doi:10.6026/97320630081230

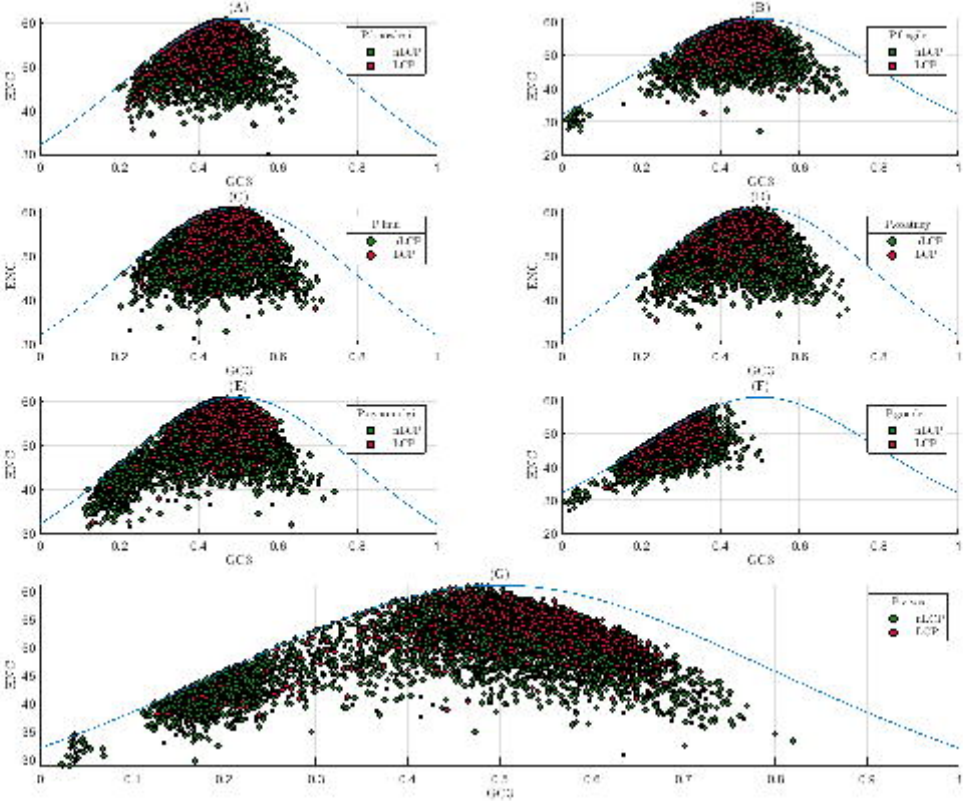

(A)
×10^4
Length
3.5
3
2.5
2
1.5
1
0.5
0

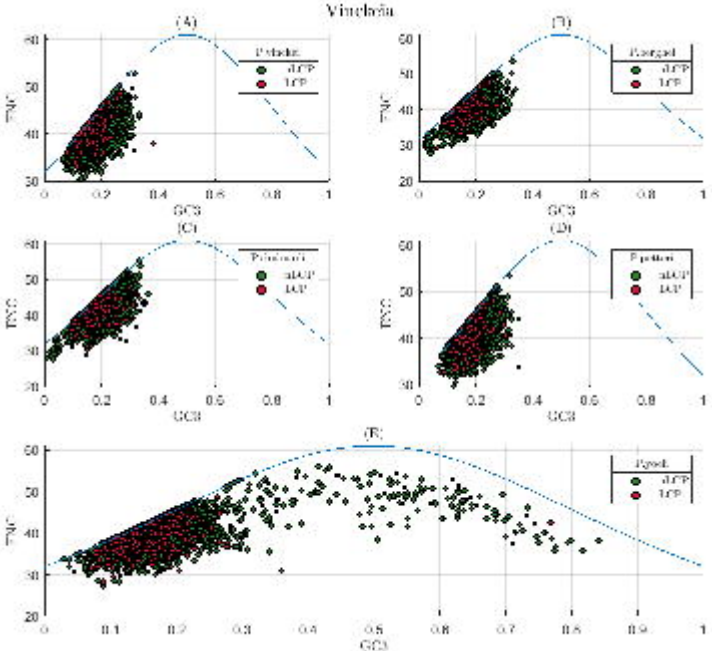

(D)
Length(LC Depleted)

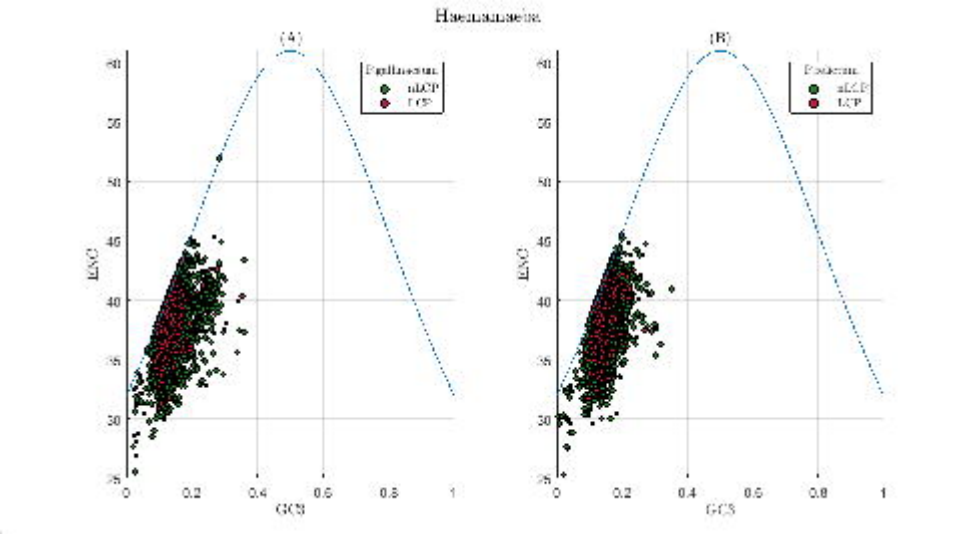
Laverania

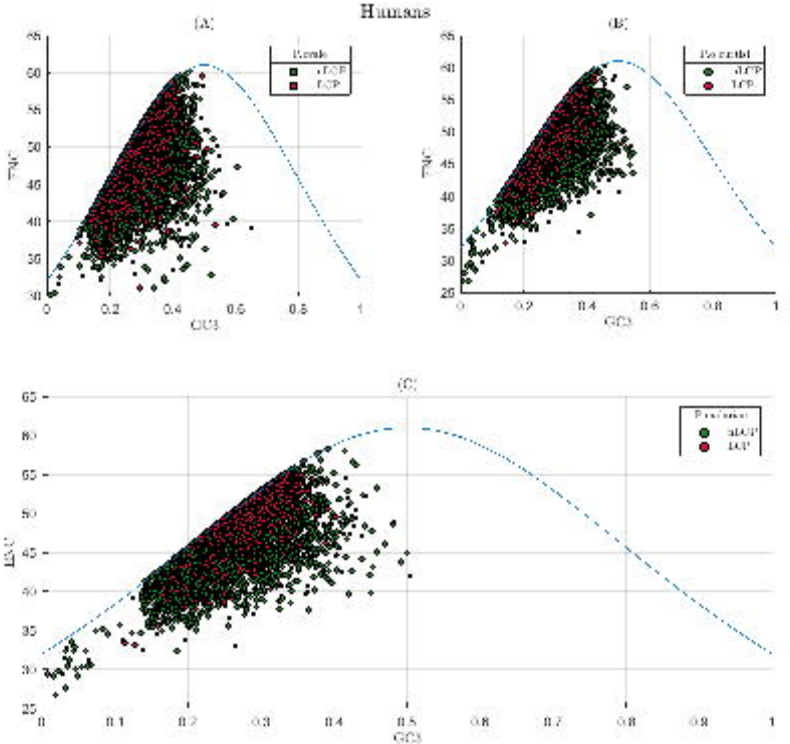


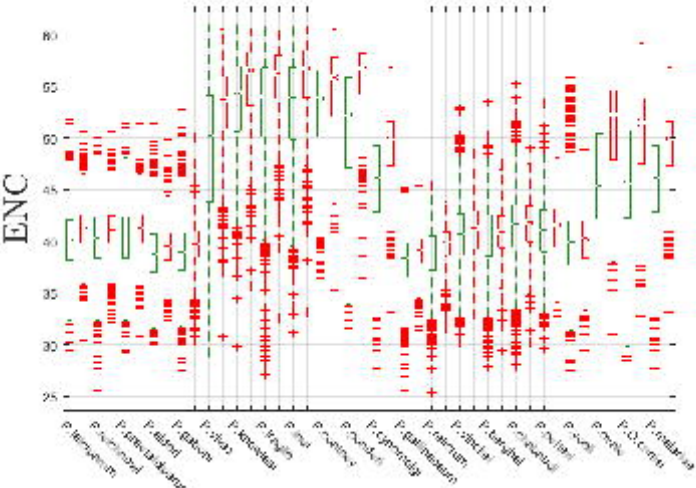
Simian Plasmodia

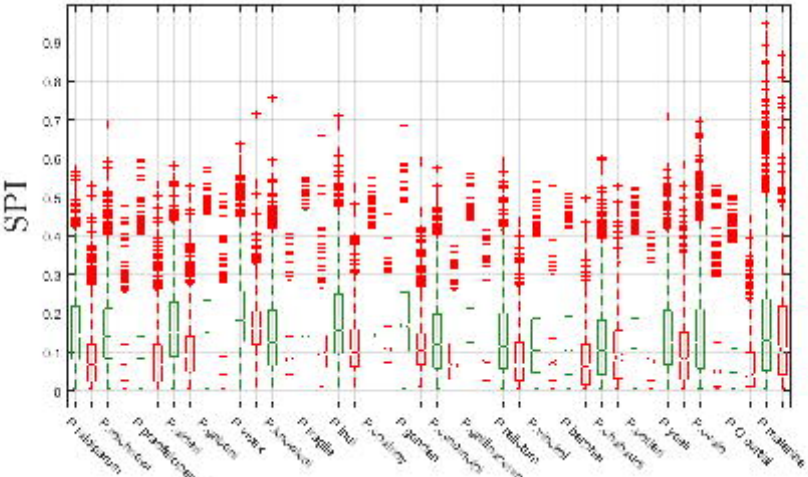



Vinclaia
(A)
(B)
(C)
(D)
(E)
TNC
GCS
dCP
LCP


Haemamaeba
(A)
(B)
Equilibrium
nLCP
LCP
ENC
GCS
0
0.2
0.4
0.6
0.8
1
25
30
35
40
45
50
55
60


Humans
(A)
(B)
(C)
GC3
GC3
GC3


ENC
80
65
60
45
40
35
30
25


SPI
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0