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:
+393493365228

Abstract

The evolutionary origin of Low Complexity Sequences (LCSs) inside the proteome of P. falciparum, the agent of the most virulent form of human malaria, is still an argument of debate. It is poorly understood whether they come from selective pressure or from mutational bias. Taking as reference several other plasmodium species and dividing each proteome in two subclasses, i.e. the Low Complexity Containing Proteins (LCPs) namely those proteins that contain at least one LCR inside their chain, and not Low Complexity Containing Proteins (nLCPs) that contain none, we find several evidences that address the LCPs as recently evolutionary emerged proteins and nLCPs as the more conserved proteins in plasmodia. The former category result to be composed by significantly longer proteins with respect to the latter. Together with the implementation of another well known bioinformatic tool The Effective Number of Codons (ENCs) and a new index that here we call the SPI (Selective Pressure Index) we find a major meddling of mutational bias on LCPs. On the other hand, the purifying pressure appears to restrain codon usage patterns in nLCPs. Our data in line with other results found in literature shows how LCPs represent an evolutionary nuance and a signal of adaptation and evolutionary path. Symmetrically, nLCPs could represent a core asset for the maintenance of parasite fitness.

Introduction

Low Complexity regions (LCRs) or non-random regions of few amino-acids are abundantly present in proteins (Kumari et al. 2005). 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 Eucarya (Wotton & Federhen, 1994).The replication slippage, a mutational process that expands or contracts microsatellites, *i.e.*, repetitive genetic sequences in which repeating units contains from one to six bases (Garcìa,2014), it has been proposed as the major factor in determining the genetic origin of LCRs (Guy-Franck et al. 2008;Guy-Franck et al. 2018;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. 2012) 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). Unfortunately, their functional history and roles are still understudied (Batistuzzi et al. 2016). P.falciparum, the etiological agent of the most virulent form of human malaria (Pizzi & Frontal 2001; Otto et al. 2018) shows a vast content of low complexity residues strongly enriched in asparagine (N) (Pizzi &Frontali,2001;Fillisetti et al. 2013;Frugier et al.2010). 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 (Fillisetti et al. 2013; Frugier et al.2010). However, different lines of evidences bring also to consider a different role for these regions. Asparagine has been found prion forming after heat shocks, that could happen after the occurrence of the febrile status of the host after the infection. This could 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 PfHsp70x prevent this aberrant aggregation (Cobb et al. 2017; Muralidharan &Golberg,2014). Muralidharan and Golberg (2014) 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 foldings and functions. Asparagine seems to represent thus a certain kind of benefit for *P. falciparum* appearing as the result of a positive selection or a luckily evolutionary accident (Muralidharan & Golberg,2014). Other results sustain that N-stretches are the outcomes also of the selection (Pizzi & Frontali, 2001) retaining that GC content could not be the unique gauge up to determine their genetic origin, while others lay the foundation of a more sophisticated mechanism proposing a multilayered action of selection, inserting the initial codon seed, and the mutational

bias weaving, by the action of replication slippage (Chaundry et al. 2017), their ubiquitous stretches among *P.falciparum* genes. 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,1986). Thus, these strands could have been inserted to respond to the needs of the parasite to be as extreme in AT content as it has been found (Pizzi &Frontali,2001; Fillisetti et al. 2013;Frugier et al.2010; Cobb et al. 2017; Muralidharan & Golberg, 2014; Chaundry et al. 2017). Since nucleotide content cannot be extremized without affecting the phenotype of a protein these strands would have been inserted, initially and then acquiring or not a classical phenotype importance, as mere placeholders to enhance AT content (Forsdyke, 2016). Finally, in line with some observations found in literature (Lipman et al.2002;Wang et al 2004) our results show how LCPs are the longest proteins in each organism here analyzed suggesting how they could be evolutionary protein nuances, while on the contrary nLCPs are the shortest proteins suggesting instead a more conservative implication for these proteins. The implementation of ENCs and SPI strengthen these preceding hypotheses catching how mutational bias acts as the priority body in shaping codon usage bias (CUB) in LCPs and how purifying selective pressure have selected a more stringent bias in nLCPs. The sum of these results suggests a crucial adaptive role for LCPs that could be representative of a signal of evolutionary path for these organisms. Conversely, ENC and SPI identify in the nLCPs a potential preserved evolutionary shell where the purifying selective pressure presses with greater force, suggesting a greater importance as regards the maintenance of the biological fitness of the parasites.

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 consider 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 by using SEG algorithm with standard parameters (Wotton Federhen,1994): W (trigger window length, an integer greater than zero, default 12 residues); K2(1) (trigger complexity, default 2.2 bits); and K2(2) (extension complexity, default 2.5 bits).

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 plasmode 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 improved implementation by Sun et al. (2012). The codon table is separated in synonymous codon families. The 6-fold codon families (Leu, Ser, and Arg) were divided into 2-fold and 4-fold codon families that can be under a dramatic different selection in CUB (Sun et al. 2012). 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. So, points that fall below the bell-shaped line suggest deviations from a null model of no-codon preferences (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 the point and Wright's theoretical curve provides an estimation of the relative extent to which natural selection and mutational bias affect codon usage of genes (Novembre, 2002). The farther is the distance, the stronger is the action of natural selection. Therefore, we propose the Selective Pressure Index (SPI), which 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 protein under study that will be characterized by a particular GC_3 content; ENCw is the ENC value of the Wright's theoretical curve for that GC_3 value. SPI helps to compare the distances under the Wright's theoretical curve being a normalized index in the range of [0, 1].We decided to implement this index because two proteins despite having the same ENC can corresponds to very different distance from the selective curve, suggesting that their CUB arose from a very different action of the natural selection. This implies that ENC cannot give a univocal information about the deviation from the null model of no-codon preference. Then, If SPI = 0 the gene will fall on the Wright theoretical curve (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 CDS sets in two branches, i.e. LCPs, with at least one LCR inside their chain and nLCPs that have none.

2.6 Cladistic and Organism Overview

The distant relation of *P. falciparum* with other human malaria parasites permitted to separate this organism in a distinct *monophyletic subgenus* termed *Laverania* (Sibley et al.2012; Liu et al. 2016).In 2009 the only other molecularly characterized plasmodium was *P.reichnowi* (Otto et al. 2018;Ollomo et al. 2009).At least six other species was added to this subgenus with three chimpanzees infecting plasmodia (*P.reichnowi, P.billcollinsi and P.gaboni*) and three that infect gorillas *(P.praefalciparum, P.blacklocki* and *P.alderi*) where we retrieved the CDS sets for *P.falciparum, P.praefalciparum, P.reichnowi, P.alderi, P.gaboni*.Among *Laveranias 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) although with not the same devasting effects. Passing from these African parassites (Otto et al. 2018) to other geographical specimens, 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 is found *P.vivax*. While *P.falciparum* it's the deadliest among the *genus 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 areas' 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;Escalante et al. 2019;Jang-Wutiwes et al. 2005). Among these *P.knowlesi* has recently emerged as a human pathogen (Rich & Xu, 2011). Thought to be strictly a primate infecting plasmodium (Faust & Dobson, 2015), naturally acquired human infections for *P.knowlesi*, in fact, were thought rare 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*, found to be related to P.cynomolgi , P.coatney and P.fragile, a naturally Bonnet macaques infecting plasmodium (Galland, 2000) and 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 this parasite (Arisue et al. 2019). Furtherly, to have a wider vision about the evolutionary, and quantitative diversification of Low Complexity Regions we considered also *rodents plasmodia* (*Vinckeia subgenus*). Among these we considered 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* (avian plasmodia) was considered, i.e. *P.gallinaceum* and *P.relictum*. Several lines of evidence propose that (Nagao et al. 2008; Waters et al. 1991; Waters et al. 1993 a-b;McCutchan et al. 1996;Escalante & Ayala,1994) P.falciparum has evolved when avians were originally domesticated therefore sharing a common ancestor (Escalante et al. 1998) with *P.gallinaceum*. Eventually we considered *P.relictum* the most geographically widespread malaria parasite for birds (Valkiunas, 2007). At last, we considered other human infecting plasmodia, 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 between plasmodia is beyond our scope. However, we found it necessary to give a brief description of the organisms taken into consideration by defining the criteria and the motivations by which they were placed in a sub-group rather than in another.

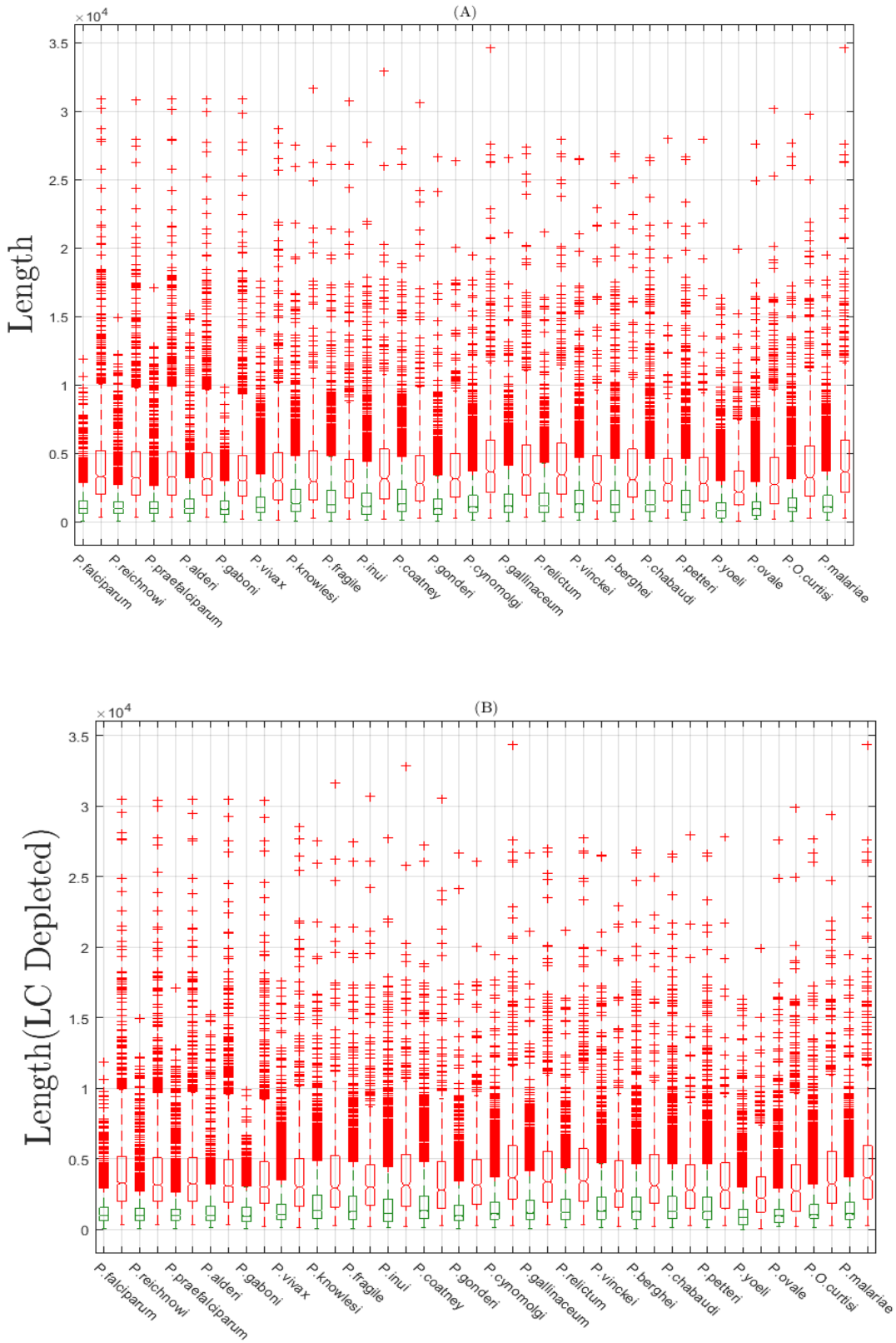


Fig 1 *Pairwise comparison for each organims for LCP(red) and nLCP(green) protein length distributions. Fig 1 (A) Protein length for LCPs is calculated with low complexity inside the protein chain. (B) Protein length for LCPs is calculted with the depletion of LCRs. Each pairwise comparison is significant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

3 Results

3.1 LCPs are Significantly Longer than nLCPs: Results and Discussion

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,1986). In order 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 LCP and nLCP protein length distributions implementing a pairwise comparison for each organism. Looking at Fig.1(A) we can see the distributions for LCPs and nLCPs, represented respectively in red and green. It can be observed that LCPs are significantly longer than nLCPs (Welch t-test; $p<<0.0001$). Furthermore implementing the *MATLAB* notch function, that permits to observe that when two notches do not overlap medians do differ with 95% confidence (Mathworks), it is possible to note 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 in each parasite we wondered if the insertion rate of the LCSs occurred in the same way for each parasite. We thus asked if the length of the LCPs was a direct consequence of the presence of the LCRs or was it a feature *de facto* (protein length favorability index for LCRs insertion or LCRs causatives of LCPs' protein length). This would have explained a possible greater probability for these stretches for squeezing into interdomains.We implemented several correlation analysis(*Pearson*) between each protein length and its LC content for each parasite collecting the results in Tab.1. Looking at the first column of Tab.1 we can observe correlation coefficients between protein length and low complexity content (all the correlations were significant $p<<0,0001$). In *Laverania* subgenus we can observe the highest correlation coefficients. It follows that the longer the protein the higher the probability for low complexity to squeeze into interdomain. Despite the length of LCPs these correlation between protein length and LC content does not appear with the same strength in all the analyzed species. This suggest that protein length is one of the possible factors that causes LC to spread among protein domains and that LCRs are not causative of the length of LCPs. This eventuality is further confirmed looking at Fig 1 (B). We depleted all LCPs of their LC content measuring the protein length as they had no LCSs natively. The pairwise comparison confirms further our precedent hypothesis namely that LCPs are *de facto* longer than nLCPs. On the other hand, in *Laverania* the low complexity content appears to be extremely higher than in the other subgenus. The second column of Tab 1 collects the average GC content of each parasite. Results emphasize how the higher the AT content the more the low complexity content inside a protein (Forsdyke, 2016) for P. falciparum. Looking in a more macroscopic way to Tab.1, this result is coherent for the whole *Laverania subgenus* and *simian plasmodia* (we considered also *P.gonderi* despite being an African monkey infecting plasmodia see Cladistic and Organism Overview). *Laverania* show at the same time a massive content of LC and a whimsical low GC content while *simian plasmodia* are characterized by a poorer content of LC residues and a less biased GC content. In fact, 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 higher GC content with a drastic reduction in LCSs content and correlation between LC insertion and protein length. On the other hand, looking at the other subgroups (*e.g. Vinckeia*) we found controversial behavior. In *Vinckeia* we do not find the same massive expression of LC residues in face of the low GC content. This eventuality is accompanied by low correlation between protein length and low complexity, that reinforces the observation made by Pizzi & Frontali (2001) who state that in P.falciparum GC content could not be the unique driving force behind the presence of LC implying that the explosion of Low Complexity could be a distinctive trait of *Laverania subgenus*. Thus, given the whimsical behavior of the *Laverania* and the low GC and LC content found in *Vinckeia* subgenus, evolutionary related to *P.falciparum* (Yadav & Swati, 2012), the GC content appear to be

not the unique responsible behind the presence of LC. Conversely, *Haemoamaeba* sees instead a closer behavior to *Laverania*. Eventually, we find *P.malariae*, *P.ovale* and *P.o.curtisi* having a GC content halfway between *simian plasmodia* and *Laverania* with a rather high LC content.

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	125557
P.reichnowi	r = 0,71	0,23	134032
P.alderi	r = 0,67	0,22	105581
P.gaboni	r = 0,70	0,22	104417
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.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	49711
P.relictum	r = 0,37	0,21	43110
P.malariae	r = 0,45	0,30	73599
P.ovale	r = 0,43	0,36	43110
P.o.curtisi	r = 0,30	0,32	32005

3.2 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 (2012). As originally formulated (Wright, 1990), the theoretical curve represents those ENC values where the synonymous codons are chosen indistinctly. CUB for these proteins must be attributed entirely to mutational bias. On the other hand, the farther the distance from the theoretical curve the higher the meddling of selective pressure in shaping codon usage bias (Novembre, 2012). To tackle the problem of Low Complexity origin in *P.falciparum*, we analyzed all the plasmodia at our disposal. Fig 2,3,4,5 and 6 show the ENC plots for *Laverania, simian plasmodia, Vinckeia, Haemamaeba,* and for the other *human infecting plasmodia* respectively. In red we represented LCPs, in green nLCPs. At first glance, LCPs show to be nearer the expected curve. This characteristic appears in all the subgroups despite the extremely wide variance of ENC values and GC_3 content. *Laverania, Vinckeia* and *Haemomaeba* show to be extremely AT rich with their ENC distributions placed on the left side of the plots. Furthermore, these organisms show a rather strong linear correlation (*Pearson*) between ENC and GC content (see Tab 2), where this relation was first spotted for *P.falciparum* in other works (Gajbhiye et al. 2017). This implies that ENC values tend to increase as the GC_3 increases as well. The same goes for *P.malariae, P.ovale* and *P.o.curtisi*. As it appears, the number of synonymous

codons utilized for these parasites tends to increase as the GC_3 content increases as well. All the correlations are significant (*p<<0,0001*). As it can be seen from Tab. 2 *Laverania* do not show the highest coefficients that instead are showed by more averaged GC containing plasmodia like *P.ovale* and *P.o.curtisi*. This correlation it is quite in pair for both nLCPs and LCPs with similar correlation coefficients in both the categories. As far as the simian plasmodia are concerned, their correlation coefficients (polynomial of second order fit) are reported in Tab.3. All the correlations are significant (*p<<0,0001*). Looking at the ENC plot all the correlations can be understood with simplicity. 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 theoretical curve that implies a greater selective pressure acting upon these.

Table 2 Correlation coefficients between GC_3 content and ENC in nLCPs and LCPs for *Laverania,Vinckeia,Haemameaba and other human infecting plasmodia*

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

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 of LCPs and nLCPs as we did for protein length distributions. Results are collected in Fig.8 (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.fragile	r-square = 0,26	r-square = 0,27
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.gonderi	r-square = 0,53	r-square = 0,59

As it can be seen LCPs in the vast majority of the cases tend to be expressed with a higher number of codons that is synonymous of a major meddling of mutational bias in shaping CUB. LCPs in *Laverania* are

expressed with greater values of ENC significantly (*p<<0,00001*) than nLCPs and similar results are found in *simian plasmodia* (*p<<0,00001*) as well as it is for *Haemameaba P.ovale,P.malariae* and *P.o.curtisi*. 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 for these species of parasites.

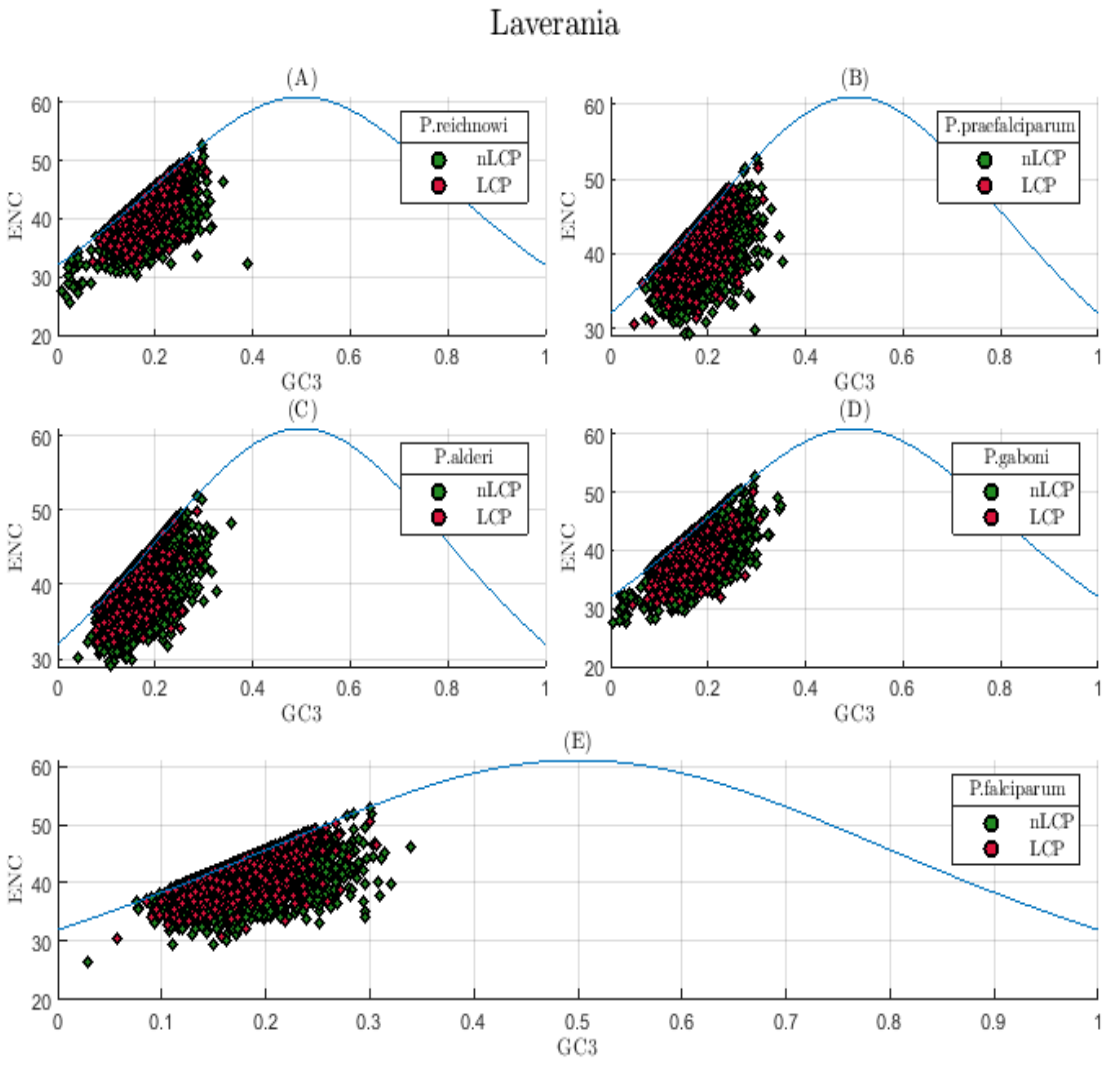


Fig 1 *ENC plot for Lavernia sub genus (A) P.reichnowi (B) P.praefalciparum (C) P.alderi (D) P.gaboni (E) P.falciparum*

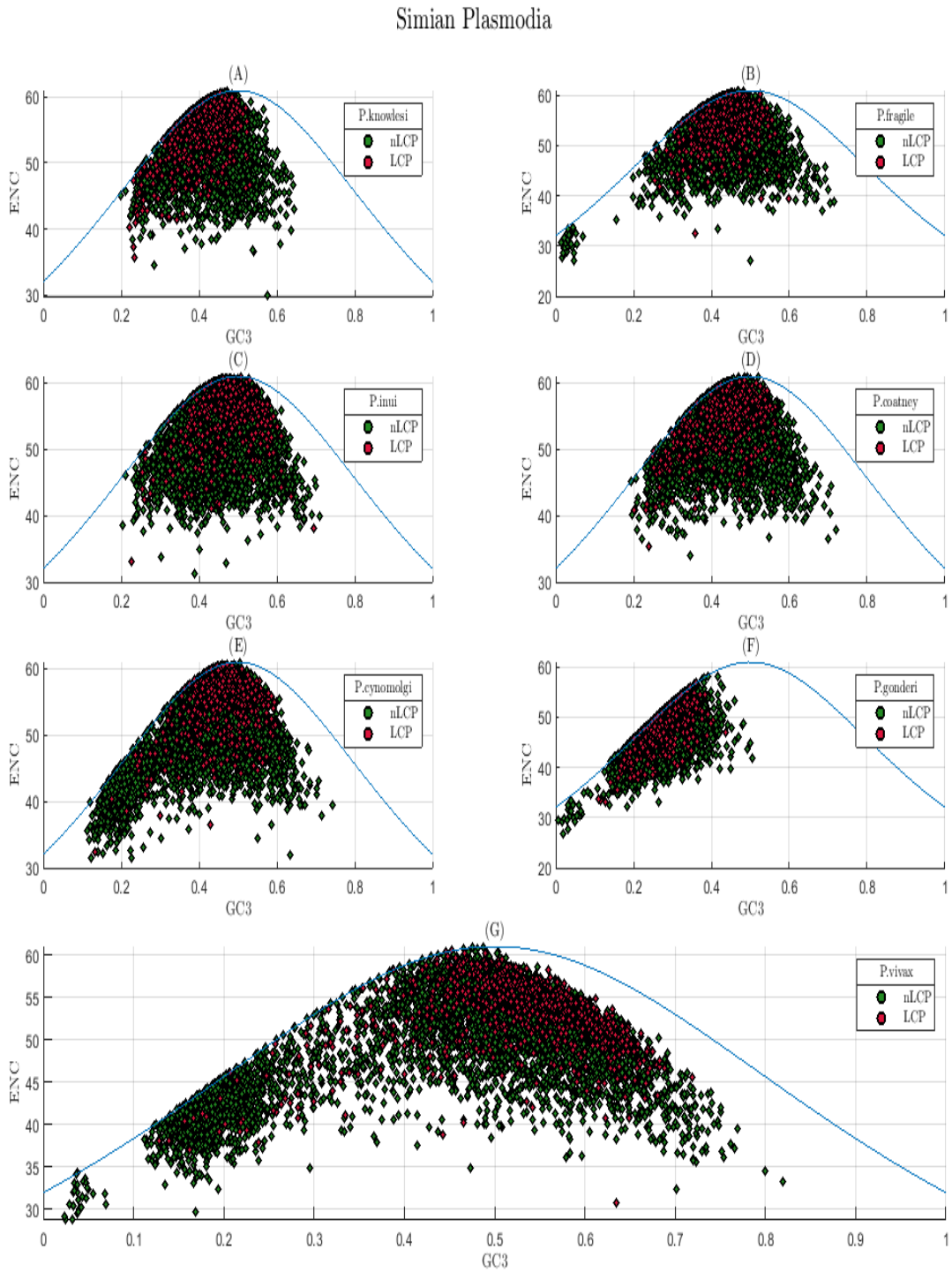


Fig 2 *ENC plot for simian plasmodia (A) P.knowlesi (B) P.fragile (C) P.inui (D) P.coatney (E) P.cynomolgi (F) P.gonderi (G) P.vivax*

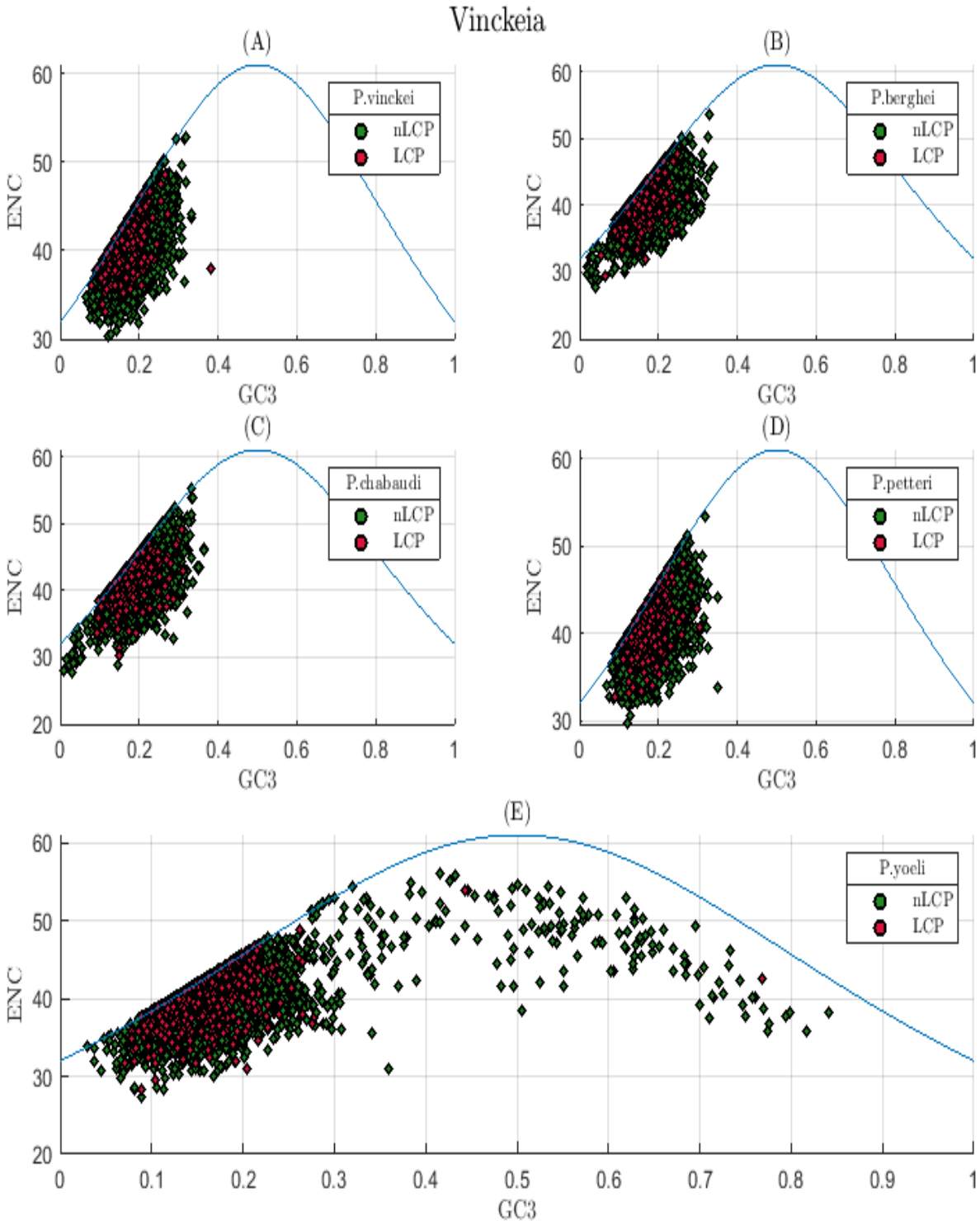


Fig 3 *ENC plot for Vinckeia subgenus (A) P.vinckei (B) P.berghei (C) P.chabaudi (D) P.petteri (E)P.yoeli*

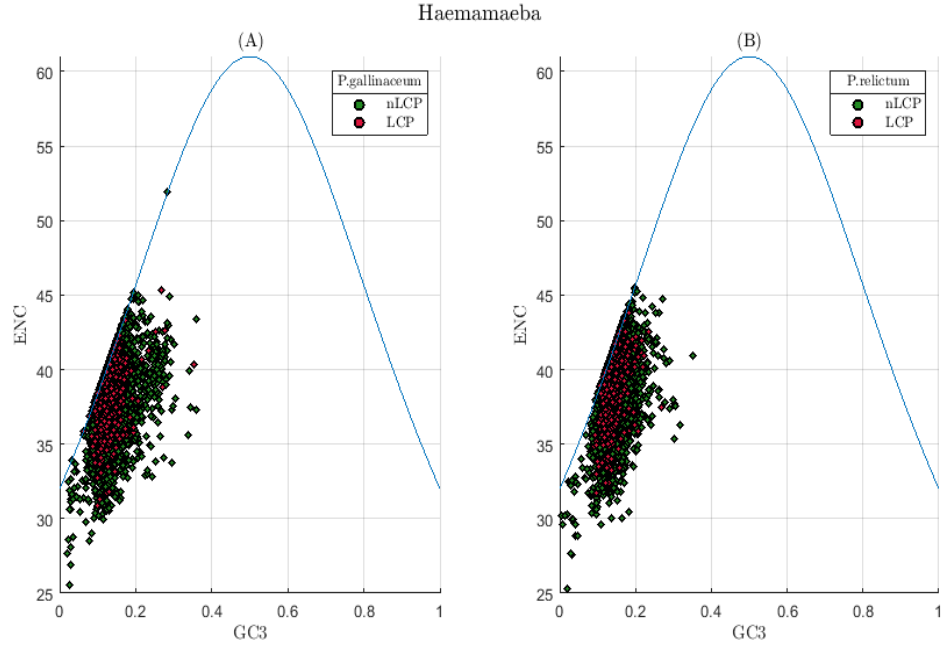


***Fig** 4 ENC plot for Haemamaeba sub genus (A) P.gallinaceum (B) P.relictum*

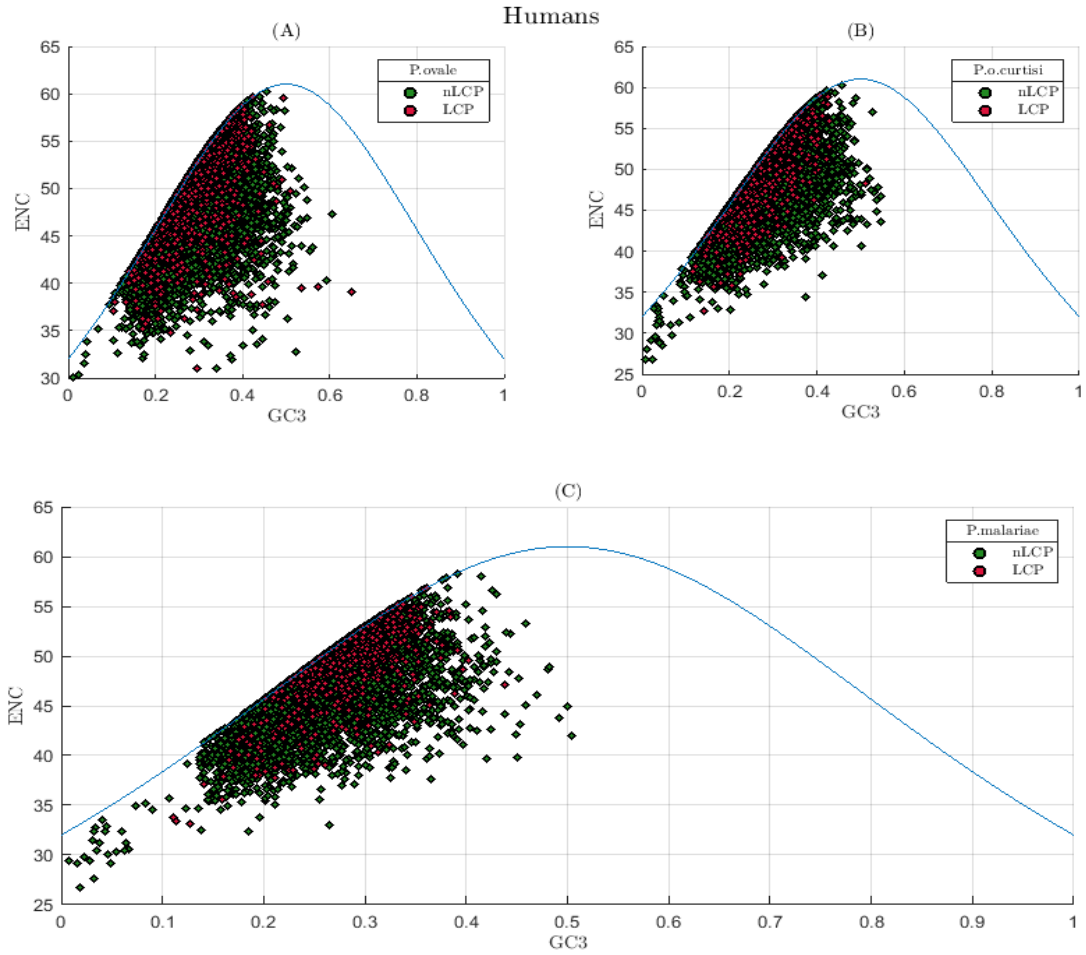


***Fig 6** ENC plot for (A) P.ovale (B) P.o.curtisi (C) P.malariae*

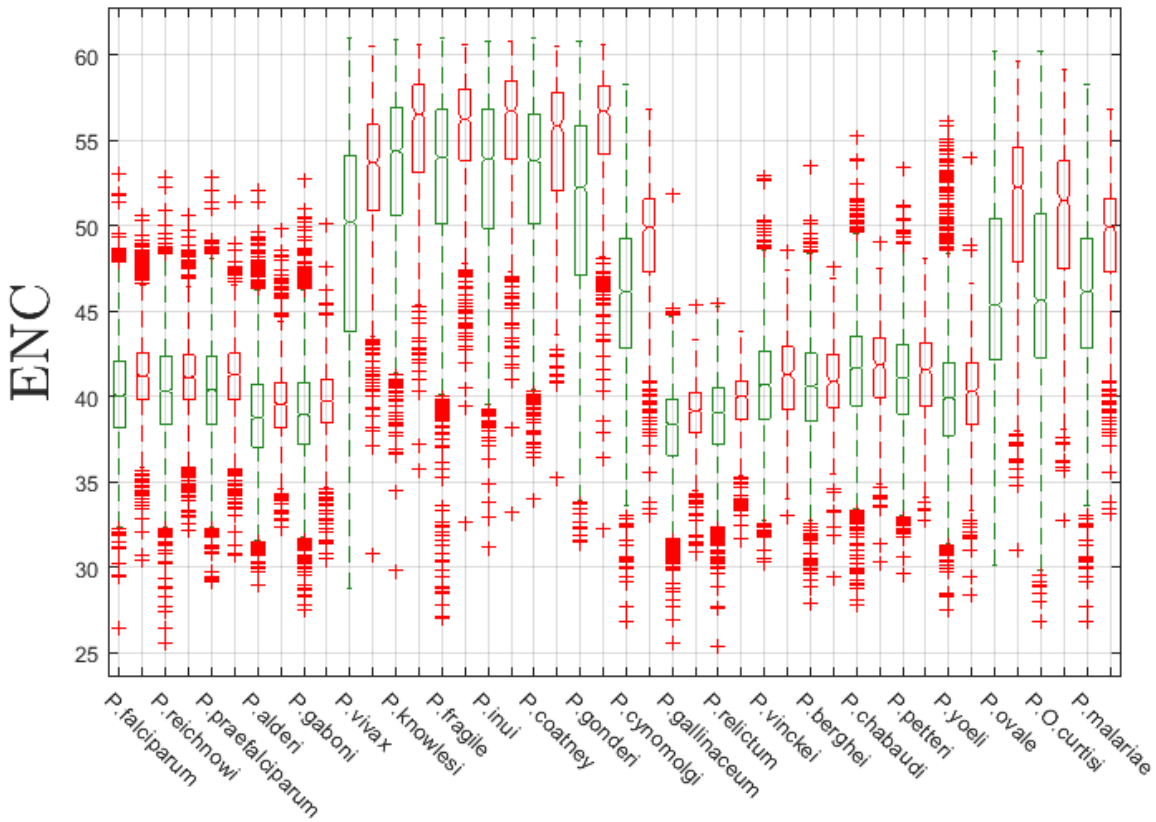


Fig 7 *ENC distributions for all the considered plasmodia. For each pair of box plots in red LCPs' ENC distribution while in green we represented nLCPs'. All the pairwise comparison are significant (p<<0,00001). However some of the vinckeia do not show these distinctions (p>0,5).*

3.3 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 CUB in LCPs. However, ENC values have some problems since the same ENC value can correspond to a very different action of 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 with former ones and displays in a better fashion the role of both mutational and selective pressure. nLCPs tend to be farther from the theoretical curve 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<<0,000001*).

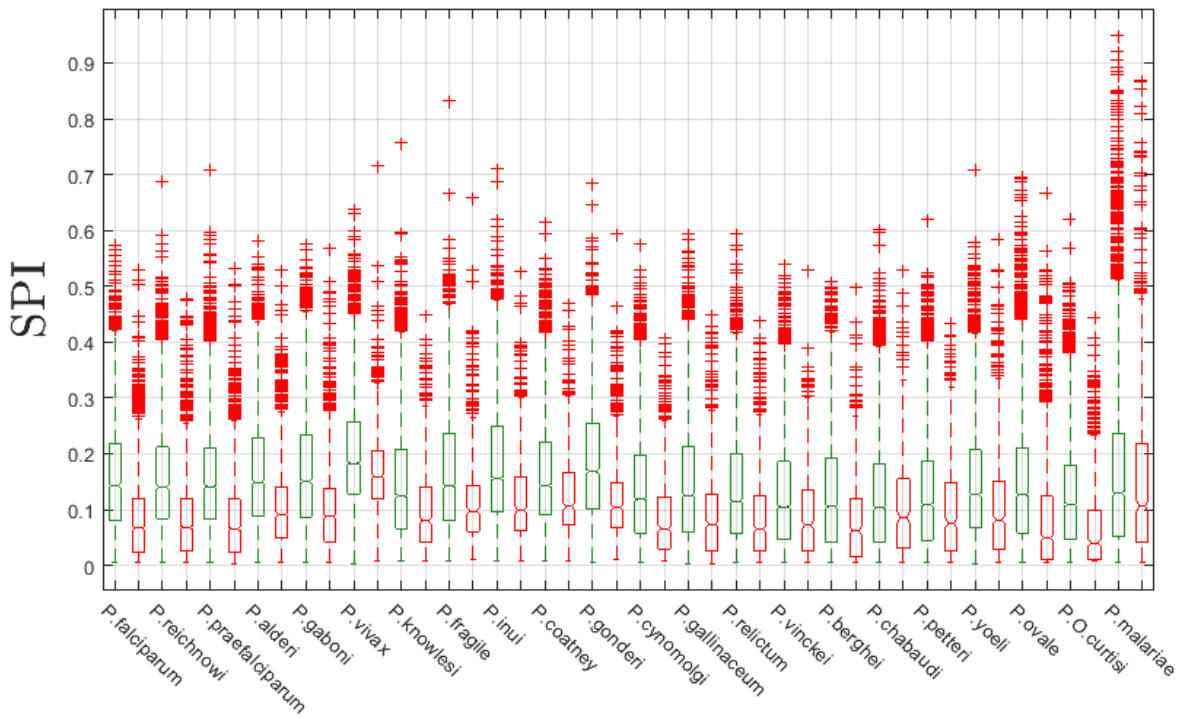


Fig 8 *SPI distributions. In red LCPs in green nLCPs. LCPs are nearer the expected curve since their SPI values tend to zero while nLCPs tend to depart more from the expected curve. All the pairwise comparisons are significant (p<<0,00001)*

Discussion

Relying on other lines of evidence found in literature and our result 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 bias in CUB would occur in highly expressed genes, while in weaker one the persistence of non-optimal codons would be allowed together with the former one due to selection for speed of translation or translational efficiency. Furthermore, selection is likely to be stronger on codons in highly expressed genes since they are used more often (Bulmer 1987;1988a;1991). 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 are the most ancient, being passed through several action of the natural selection that finally selected those patterns of codons that allowed the organism to survival, to reproduce optimizing its internal biosynthetic ecosystem. Indeed, results see how evolution would reward shorter proteins (Lipman et al. 2002) while longer proteins resulted to be, on average, evolutionary nuances (Wang et al. 2004) appearing to have emerged in more recent times. Protein elongation can occur in many ways such as domain duplication or insertion believed to be a general tendency in protein evolution (Wang et al 2004). Furthermore, newly evolved proteins show to have a greater content in LCRs (Toll-Riera et al. 2012) where primordial proteins are hypothesised to be highly repetitive oligomers that specified periodical polypeptides (Ohno & Epplen,1983). These repetitive sequences should 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. On the other hand, LCRs are found in performing some functions (Shen et al.2004; Salichs et al.2009). Thus, it results to be logical, in some cases, to think to a certain importance in maintaining the fitness of the organism that could imply a major role of the selective

pressure in preserving these regions in particular cases. Otherwise they can also vanish (Pancsa & Tompa, 2016). Our main results can be summarised in this way: nLCPs are (i) shorter than LCPs. Low Complexity stretches does not squeeze into interdomain equally in all the organism specific LCPs. This enlighten that LCPs are *de facto* longer than nLCPs, characteristic that provide a fertile field for Low Complexity to be inserted along the protein chain. Furthermore, the absence of a uniform presence of LC in relation of GC content underline that in *P.falciparum* and more widely in *Laverania* the ubiquitous presence of LC does not depend from GC only. To know if this lack of complexity is conserved also at mRNA level is left to other works. nLCPs (ii) are characterized by a more stringent bias in Codon Usage while LCPs see a more relaxed action of purifying selection given the presence of a wider plethora of used codons. Given unclear situations found in *Vinckeia* that sees a poor distinction in CUB between nLCPs and LCPs (iii) SPI solves these problems where LCPs witness a minor depart from the expected curve. The puzzling situation found in *Vinckeia* is to address to similar ENC and different GC_3 values (see Methods) that can puzzle when it is to evaluate the role of selection and the mutational noise representaned by mutational pressure.the Furthermore, we found that the longer the LCP the more the Low Complexity Residues expressed in its chain. The sum of our results in reference to other literature works (Toll-Riera et al. 2012; Lipman et al. 2002; Wang et al. 2004) brings us to attribute to LCPs and nLCPs different evolutive roles and importance. nLCPs appear to be under a more stringent action of selective pressure suggesting a difficulty to accept mutations that could impair the normal function and folding on which the parasites build its metabolism. The presents appear to be crucial for the parasite survival. On the other hand, the characteristics of LCPs underline a lack of structural definition implying where selection could slice and dice with more freedom without impairing necessarily the fitness of these organisms providing a fertile ground for proteins to acquire new folds and functions potentially providing adaptive strategies for the *plasmodia* to evolve and differentiate. The less marked action of selective pressure implies the capacity of LCPs to acquire mutations without impairing their functions. In this sense LCPs seem to be recently evolved proteins that are “still searching” for a molecular destiny. In this sense LCRs would serve as a black sheet where selection could design new structural architectures gradually acquiring mutations bringing the parasite to adapt and evolve. Is still puzzling the role of Replication slippage given the great diversity in Low Complexity presence among these parasites. Our position sustains a precedingly statement made by Chaundry and colleagues (Chaundry et al. 2017) which suggest that both the selection and mutational bias drive the insertion of Asparagine in LC domains in *P.falciparum*. More studies must be done, first a further analysis by Neutrality and Pr2 (Sueoka,1995,1998,1999) plot to deepen the knowledge about evolutionary pressures. The analysis of Low complexity codon content alone is necessary to understand if the replication slippage is really a strong phenomenon. In fact, the Low Complexity could not be conserved at mRNA level and this would impair the mechanism of the replication slippage (Gemayel et al. 2012). In such a way, it will be possible to understand if the GC content affected LCRs giving a further explanation in regard of the whimsical LC content in *Laverania*. On this line an analysis of RSCU (Sharp & Li,1987) on these stretches could answer solving this riddle.

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 disordered proteins as reservoirs for evolutionary innovations and nLCPs as the core asset for parasites’ survival. 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. Moreover, extending the study

presented here to other *apicomplexa* could add further comparative insight into the repertoire of the adaptive evolutionary mechanisms of protein.

References

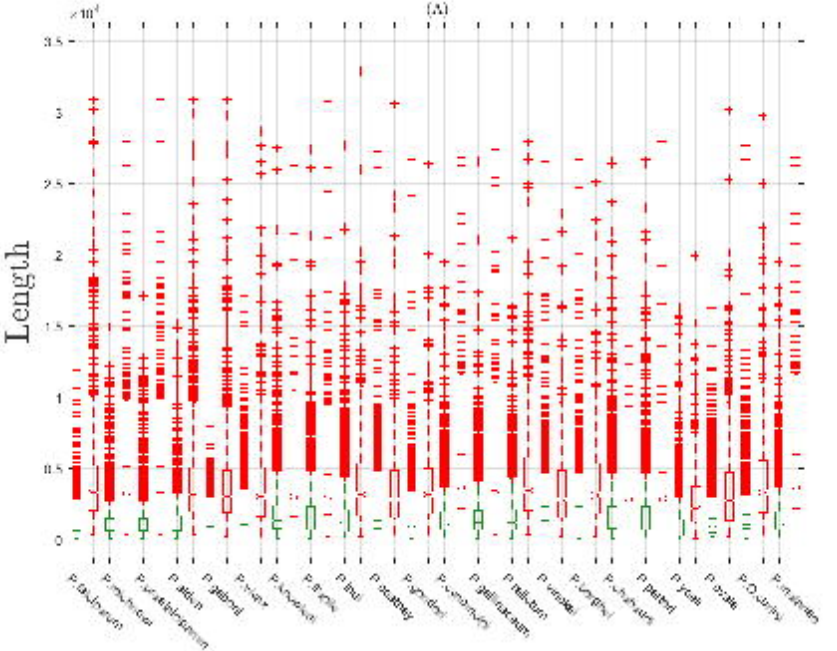
- Akashi H, Translational selection and yeast proteome evolution. Genetics 164: 1291-1303
- Arisue, N., Hashimoto, T., Kawai, S. et al. Apicoplast phylogeny reveals the position of Plasmodium vivax basal to the Asian primate malaria parasite clade. Sci Rep 9, 7274 (2019). https://doi.org/10.1038/s41598-019-43831-1
- 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
- Belalov IS, Lukashev AN (2013). Causes and implications of codon usage bias in RNA viruses. PLoS ONE 8:e56642
- Brown LY, Brown SA. Alanine tracts: the expanding story of human illness and trinucleotide repeats, Trends Genet., 2004, vol. 20 (pg. 51-58)
- 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
- Cobb DW, Florentin A, Fierro MA, Krakowiak M, Moore JM, Muralidharan V. The Exported Chaperone PfHsp70x Is Dispensable for the Plasmodium falciparum Intraerythrocytic Life Cycle. mSphere. 2017;2(5):e00363-17. Published 2017 Sep 27. doi:10.1128/mSphere.00363-17
- 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
- Ellegren H. Microsatellites: simple sequences with complex evolution, Nat Rev Genet. , 2004, vol. 5 (pg. 435-445)
- Escalante A A, Ayala F J. Proc Natl Acad Sci USA. 1994;91:11373–11377
- Escalante AA, Pacheco MA. Malaria Molecular Epidemiology: An Evolutionary Genetics Perspective. *Microbiol Spectr*. 2019;7(4):10.1128/microbiolspec.AME-0010-2019. doi:10.1128/microbiolspec.AME-0010-2019
- Escalante Ananias A., Freeland Denise E., Collins William E., Lal Altaf A., 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 Jul 1998, 95 (14) 8124-8129; DOI: 10.1073/pnas.95.14.8124
- 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 C, Dobson AP. Primate malarias: Diversity, distribution and insights for zoonotic *Plasmodium*. *One Health*. 2015;1:66□75. Published 2015 Oct 24. doi:10.1016/j.onehlt.2015.10.001
- Faust, Christina & Dobson, Andrew. (2015). Primate malarias: Diversity, distribution and insights for zoonotic Plasmodium. One Health. 1. 10.1016/j.onehlt.2015.10.001.

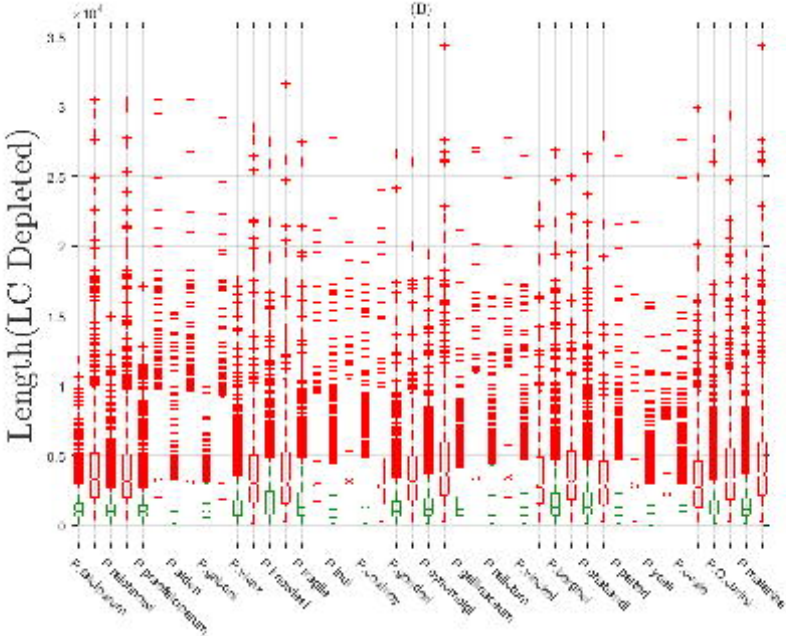
- Filisetti, D., Th´eobald-Dietrich, A., Mahmoudi, N., Rudinger-Thirion, J., Candolfi, E., Frugier, M. (2013). Aminoacylation of Plasmodium falciparum tRNA Asn and Insights in the Synthesis of Asparagine Repeats. Journal of Biological Chemistry, 288(51), 36361–36371. doi:10.1074/jbc.m113 .522896
- Forcelloni, 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
- Fuglsang A (2006) Accounting for background nucleotide composition when measuring codon usage bias: brilliant idea, difficult in practice. Mol Biol Evol 23:1345–1347
- G. Gale Galland, 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
- 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. Gale (2000). "Role of the Squirrel Monkey in Parasitic Disease Research". ILAR Journal. 41: 37–43. doi:10.1093/ilar.41.1.37. PMID 11421222. Retrieved 21 February 2018.
- 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
- Guy-Franck R. & Mosbach, Valentine & Poggi, Lucie. (2018). Trinucleotide repeat instability during double-strand break repair: from mechanisms to gene therapy. Current Genetics. 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.
- 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)
- José Marín-García M.D., in Post-Genomic Cardiology (Second Edition), 2014
- 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
- 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
- Moriyama, E. N., and J. R. Powell. 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. 26:3188–3193.
- 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
- Nagao E, Arie T, Dorward DW, Fairhurst RM, Dvorak JA. The avian malaria parasite Plasmodium gallinaceum causes marked structural changes on the surface of its host erythrocyte. *J Struct Biol*. 2008;162(3):460□467. doi:10.1016/j.jsb.2008.03.005
- 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. (1984). Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proceedings of the National Academy of Sciences, 81(8), 2421–2425. doi:10.1073/pnas.81 .8.2421
- 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
- Ollomo B, et al. A New Malaria Agent in African Hominids. PLoS Pathog. 2009;5:e1000446. doi: 10.1371/journal.ppat.1000446
- 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
- Pal, C., B. Papp, and L. D. Hurst. 2001. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol. Biol. Evol. 18:2323–2326
- Pancsa R., Tompa P., Coding Regions of Intrinsic Disorder Accommodate Parallel Functions, Published:September 16, 2016DOI: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

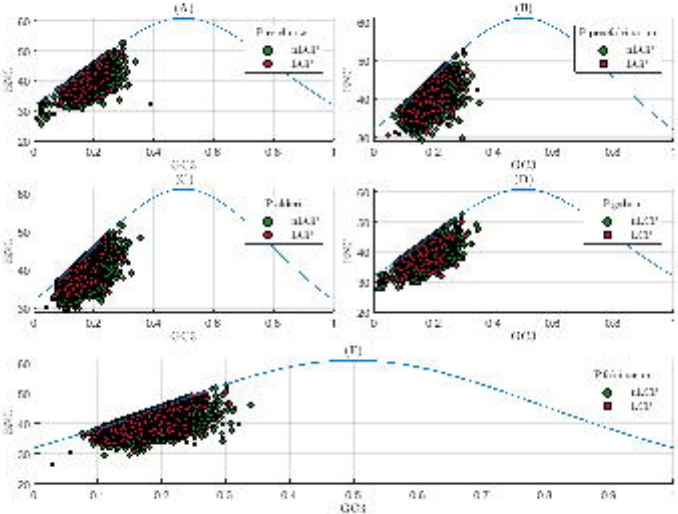
- Prugnolle F, Durand P, Neel C, et al. African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. *Proc Natl Acad Sci U S A*. 2010;107(4):1458□1463. doi:10.1073/pnas.0914440107
- Rich SM, Xu G. Resolving the phylogeny of malaria parasites. Proc Natl Acad 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):1281□1295. 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
- Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657
- Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325
- Sueoka N (1999) Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. Gene 238:53–58
- 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 Wootton
- 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.
- Verzier LH, Coyle R, Singh S, Sanderson T, Rayner JC (2019) Plasmodium knowlesi as a model system for characterising Plasmodium vivax drug resistance candidate genes. PLoS Negl Trop Dis 13(6): e0007470. https://doi.org/10.1371/journal.pntd.0007470
- 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

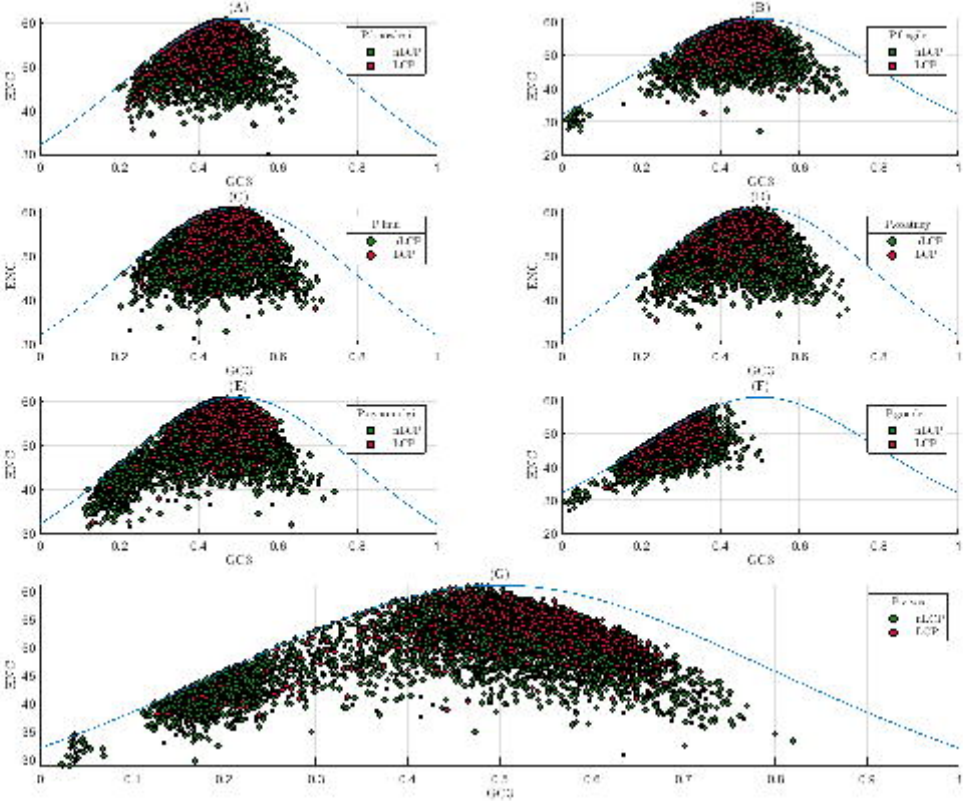

(A)
×10^4
Length
3.5
3
2.5
2
1.5
1
0.5
0

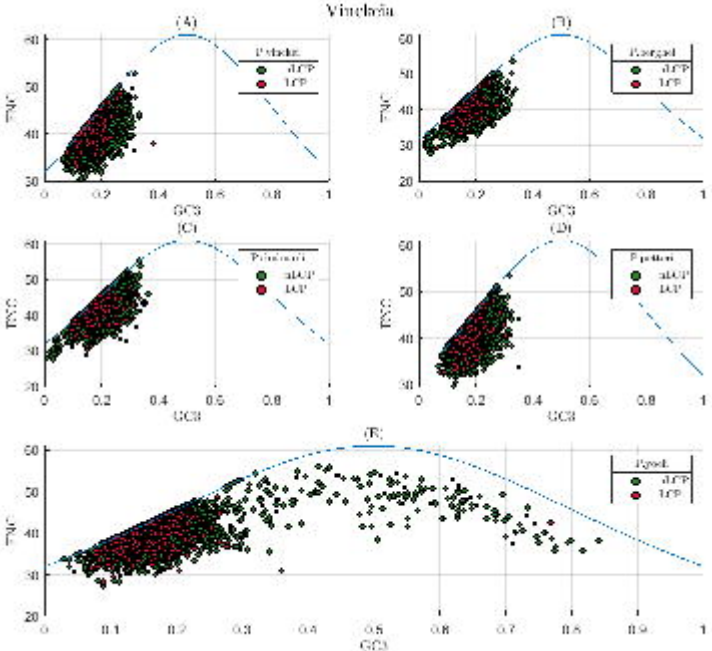

(D)
Length(LC Depleted)

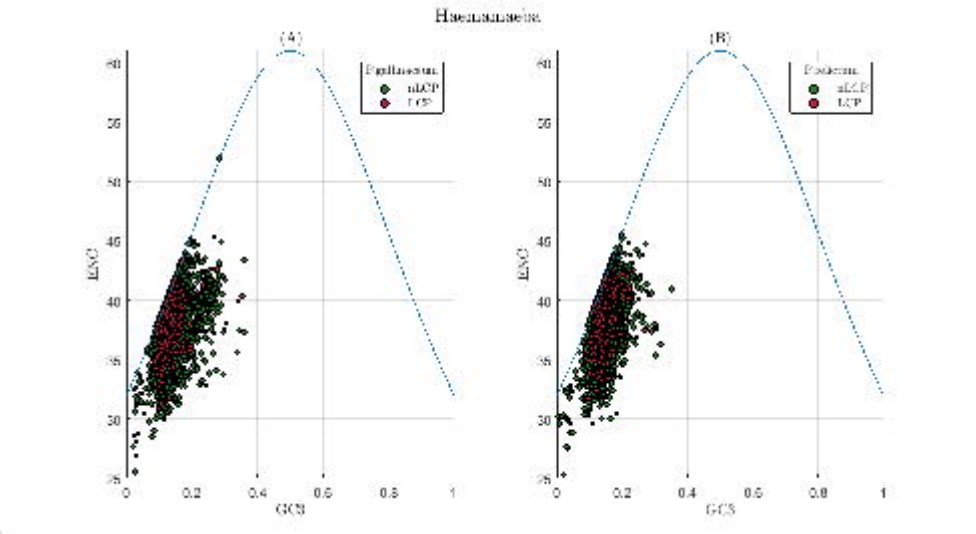
Laverania

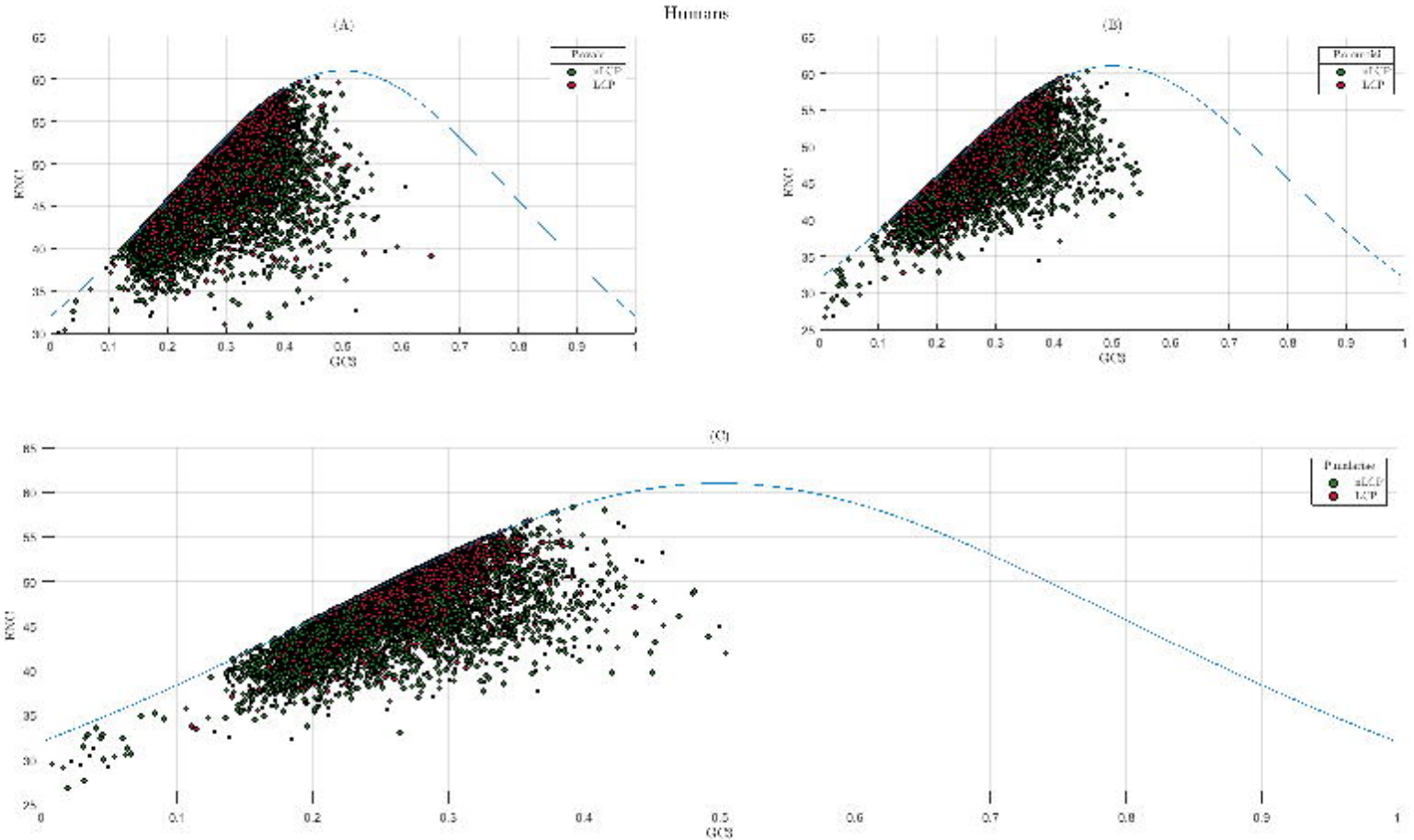


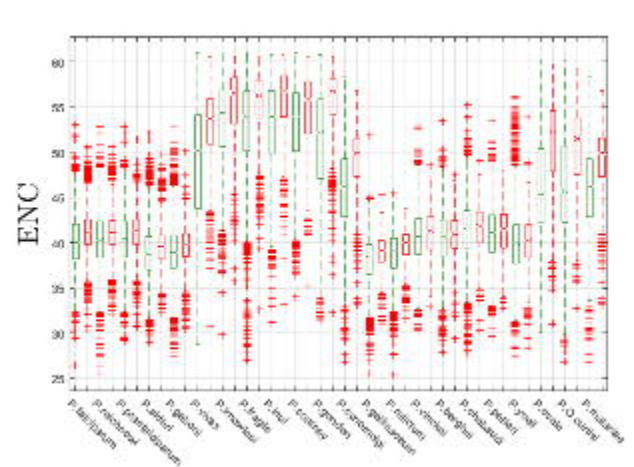
Simian Plasmodia

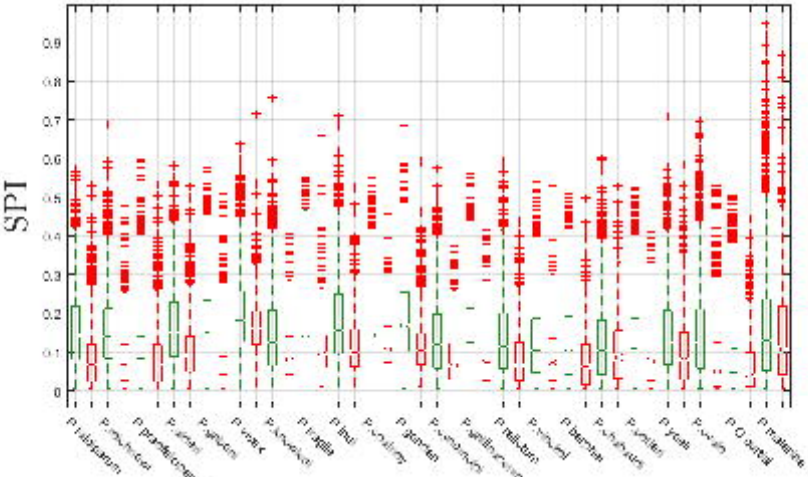



Vinclaia
(A)
(B)
(C)
(D)
(E)
TNC
GCS
dCP
LCP


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)
KNC
GCS
aLCP
LCP


ENC
60
55
50
45
40
35
30
25
P. falciformis
P. roborescens
P. pseudolapidarius
P. atlicolor
P. geotropi
P. vaces
P. ostreatus
P. fragile
P. inuli
P. cornucopiae
P. gordonii
P. cornucopiae
P. gallinaceus
P. nidulans
P. vinicolor
P. chrysoloma
P. petaloides
P. pinsitus
P. ostreatus
P. O. cornucopiae
P. nidulans


SPI
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0