Πανεπιστήμιο Κύπρου

**Department of Biological Sciences**

# COMPOSITIONALLY BIASED REGIONS: FROM *STRUCTURAL SIGNATURES* TO *COMPARATIVE GENOMICS*

**TAMANA STELLA**

**A dissertation submitted to the University of Cyprus in partial fulfillment of the requirements for the degree of Doctor of Philosophy**



**December 2018**

**VALIDATION PAGE**

**Doctoral Candidate: Stella Tamana**

**Doctoral Thesis Title:** Compositionally Biased Proteins: from structural signatures to comparative genomics

*The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Philosophy at the* ***Department of Biological Sciences*** *and was approved on the 6th of December by the members of the* ***Examination Committee****.*

**Examination Committee:**

**Research Supervisor: Dr. Vasilis Promponas, Assistant Professor**

(Name, position and signature)

**Committee Member: Dr. Pantelis Georgiades, Assistant Professor**

(Name, position and signature)

**Committee Member: Dr. Antonis Kirmizis, Associated Professor**

(Name, position and signature)

**Committee Member: Dr. Gregorios Amoutzias, Assistant Professor**

(Name, position and signature)

**Committee Member: Dr. Georgios Pavlopoulos, Associated Professor**

(Name, position and signature)

## DECLARATION OF DOCTORAL CANDIDATE

The present doctoral dissertation was submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Stella Tamana

……

**ABSTRACT** [in greek language]

Οι Περιοχές Ακραίας αμινοξικής Σύστασης (ΠΑΣ) αναφέρονται σε τμήματα κατά μήκος πρωτεϊνικών αλληλουχιών, όπου η σύστασή τους ευνοεί ένα/λίγα αμινοξικά κατάλοιπα. Οι ΠΑΣ βρίσκονται σε αφθονία σε όλους τους οργανισμούς και συχνά συσχετίζονται άμεσα με συγκεκριμένα δομικά πρότυπα, με εμφανείς συνέπειες στη (δυσ)λειτουργία των πρωτεϊνών. Επιπλέον, οι αιτιολογικοί παράγοντες της μαλάριας, τα παρασιτικά πρωτόζωα του γένους *Plasmodium*, παρουσιάζουν ασυνήθιστα υψηλή συχνότητα εμφάνισης ΠΑΣ, λόγω των εμπλουτισμένων γονιδιωμάτων τους σε Αδενίνη και Θυμίνη. Τα εξαιρετικά εμπλουτισμένα γονιδιώματα των παρασίτων της μαλάριας προκαλούν επιπρόσθετες τεχνικές προκλήσεις για τον προσδιορισμό της αλληλουχίας του γονιδιώματος, την κλωνοποίηση σε συστήματα ετερόλογων φορέων και απαιτούν ειδική μεταχείριση σε βασικά στάδια της συγκριτικής γονιδιωματικής. Επιπρόσθετα, ο περίπλοκος κύκλος ζωής των Πλασμωδίων ταξινομεί τα είδη αυτά ως εξαιρετικά πολύπλοκους οργανισμούς για την διεξοδική μελέτη της βιολογίας τους *in vivo*. Ωστόσο, ο αυξημένος αριθμός των Πλασμωδίων με πλήρη αλληλουχία γονιδιώματος, τα καθιστά σαν ιδανικό είδος για υπολογιστικές μελέτες σχετικά με το ρόλο των ΠΑΣ στην παθογένεια και την εξελικτική συμπεριφορά αυτών και άλλων ειδών με ΠΑΣ.

Η παρούσα μελέτη περιστρέφεται γύρω από τρία βασικά χαρακτηριστικά των ΠΑΣ: (i) πως οι ΠΑΣ επηρεάζουν τους υπολογισμούς κατά την ανάλυση παν-γονιδιωμάτων, (ii) τον ρόλο των ΠΑΣ στην εξελικτική συμπεριφορά των εξεταζόμενων πρωτεϊνικών οικογενειών, και (iii) τις δομικές υπογραφές των ΠΑΣ. Για την διεξαγωγή της παρούσας μελέτης χρησιμοποιήσαμε σύνολα δεδομένων που ανακτήθηκαν από ειδικευμένες βάσεις δεδομένων και αναπτύξαμε νέες υπολογιστικές διαδικασίες για την απάντηση των παραπάνω ερωτημάτων.

Τα αποτελέσματα μας υποδεικνύουν ότι η προκύπτουσα δομή του παν-γονιδιώματος των Πλασμωδίων εξαρτάται σε μεγάλο βαθμό από τις διαφορετικές στρατηγικές που χρησιμοποιήσαμε για την ανίχνευση των ΠΑΣ. Επιπρόσθετα, βασιζόμενοι τόσο σε στατιστική αλλά και βιολογικής σημαντικότητας ανάλυση, προτείνουμε μια βέλτιστη στρατηγική για συγκριτική γονιδιωματική παν-γονιδιωμάτων με υψηλό περιεχόμενο ΠΑΣ.

Κατά την παραπάνω συγκριτική γονιδιωματική ανάλυση, ανακαλύψαμε ευρήματα που αμφισβητούν την ευρέως διαδεδομένη αντίληψη ότι μόνο τέσσερις από τις οχτώ υπό-μονάδες του συμπλόκου OST —που θεωρείται διατηρημένο σε όλους τους ευκαρυωτικούς οργανισμούς —είναι παρούσες στα Πλασμώδια και άλλα πρώτιστα. Είναι αξιοσημείωτο, ότι ο κύριος λόγος για τον οποίο η ασυνήθιστα βραχεία πρωτεΐνη Ost4 (36 κατάλοιπα στους ζυμομύκητες) δεν είχε εντοπιστεί μέχρι στιγμής, είναι η αποτυχία των υπολογιστικών μεθόδων πρόβλεψης γονιδίων για την ανίχνευση μιας τέτοιας μικρής κωδικής αλληλουχίας. Στην πραγματικότητα, έχουμε σαφή ένδειξη με βάση ομοιότητες αλληλουχιών αλλά και επιπλέον υποστήριξη από δεδομένα έκφρασης (EST, RNAseq) ότι όλες οι υπό-μονάδες του συμπλέγματος OST (με εξαίρεση την υπό-μονάδα Swp1/Ribophorin II), μπορούν να ταυτοποιηθούν με αξιόπιστο τρόπο στα Πλασμώδια, αλλάζοντας την καθιερωμένη αντίληψη για την εξέλιξη του συμπλόκου OST και ανοίγοντας νέες κατευθύνσεις για την κατανόηση της πιθανής βιολογικής σημασίας του μηχανισμού της N-γλυκοζυλίωσης στα πρωτόζωα.

Παράλληλα, χρησιμοποιήσαμε αντίστροφη μηχανική και κατηγοριοποίηση για την χαρτογράφηση της αλληλουχίας των ΠΑΣ σε πειραματικά λυμένες πρωτεϊνικές δομές. Εστιάσαμε σε ΠΑΣ πλούσιες σε γλουταμικό και ασπαρτικό οξύ, όπου επανειλημμένα ομαδοποιούνται με βάση ανεξάρτητα κριτήρια, συζητώντας πιθανούς βιολογικούς τους ρόλους. Ενώ οι περισσότερες πρωτεϊνικές δομές κατατάσσονται λειτουργικά είτε ως πρωτεΐνες μεταφοράς ή ένζυμα, παρατηρήθηκαν επίσης ορισμένες επιπλέον κατατάξεις, όπως τοξίνες, πρωτεΐνες της απόπτωσης και κυτταρικής προσκόλλησης.

**ABSTRACT** [in an international language]

Compositionally biased regions (CBRs) refer to spans along protein sequences with composition favoring one or a few residue types. CBRs are ubiquitous and are quite often directly related to specific structural patterns, with apparent implications in protein (dys-)function and interactions.

*Plasmodium* species, the causative agents of malaria, have an unusual high incidence of CBRs owing to their enriched A+T genomes. The extremely biased genomes of malaria parasites induce additional technical challenges for genome sequencing projects, cloning in heterologous vector systems and require special treatment in fundamental steps of comparative genomic analyses. Additionally, the complicate life cycle of *Plasmodium* species classifies these species as very complex organisms to thoroughly study their biology *in vivo*. However, the increased number of fully sequenced *Plasmodium* genomes makes it the ideal taxon for *in silico* studies of the role of CBRs in pathogenicity, their evolutionary behavior and to evaluate strategies for handling CBRs in pan-genome analysis in the present of extreme CBR-content.

The current thesis revolves around three main aspects of CBRs: (i) how CBRs affect the computation of heavily biased pan-genomes, (ii) the role of CBRs in evolutionary behavior of the protein families under study, and (iii) CBR preferences in structural conformations. We used carefully compiled datasets of *Plasmodium* species and non-redundant protein chains retrieved from the PlasmoDB and the Protein Data Bank. Furthermore, we developed novel computational pipelines for CBR detection and masking, sequence comparison/clustering, as well as structural feature computation.

Our results indicate that our view of the plasmodial pan-genome structure is largely dependent on the different strategies used to handle CBRs. We further propose an optimal strategy for comparative genomic analyses based on thorough statistical and biological assessment.

Our comparative genomics data led us to challenge the currently established notion that only 4 out of 8 subunits of the oligosaccharyltransferase (OST) complex –which is considered conserved across eukaryotes– are present in *Plasmodium* species and other protists. Remarkably, the main reason why the unusually short Ost4 protein (36 amino acid residues in yeast) has not been identified so far is the failure of gene-

prediction pipelines to detect such a short coding sequence. In fact, based on carefully conducted sequence comparisons we provide unequivocal evidence that all components of the OST complex, with the exception of Swp1/Ribophorin II, can be reliably identified within completely sequenced plasmodial genomes. Importantly our findings are further supported by publicly available EST and RNAseq expression data.

Finally, a reverse engineering approach was followed for mapping sequence and structural signatures of CBRs in experimentally 3D solved protein structures. Specifically, our results portrayed both the structural preferences of CBRs and the sequence features these CBRs possess. We focus on D-/E-rich CBRs, which consistently are sub-clustered across the structure-based clustering, discussing their possible biological roles. Most such protein structures are classified as protein binding, transport proteins or enzymes; notably, some additional classifications, such as toxins, apoptosis and cell adhesion, were also observed.

**Acknowledgements**

*"No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude."*

Alfred North Whitehead

*Firstly, I would like to express my sincere gratitude for my advisor Dr. Vasilis J. Promponas for his precious help and guidance for the development of the present thesis until its completion. For his patience, motivation, immense knowledge and confidence, I could not have imagined having a better advisor and mentor for my PhD study.*

*Beside my advisor, I would like to thank the members of the examination committee Dr. Pantelis Georgiades, Dr. Antonis Kirmizis, Dr. Georgios Pavlopoulos and Dr. Gregorios Amoutzias for their time and evaluating the current thesis.*

*A very special gratitude goes to my colleagues and friends from the Bioinformatics Research Laboratory, Dr. Ioannis Kirmitzoglou, Dr. Athina Theodosiou, Dr. Ioanna Kalvari, Andreas-Nicolas Ioannides and Dr. Kalliopi Georgiades for their contribution to overcome several puzzling issues that emerged, stimulating discussions and for all the fun we had.*

*Also, I wish to thank Sophia Mathioudakis for her support and encouragement that I will (someday) complete this research.*

*Last but by no means the least, I would like to thank my family: my mother and in-laws, my beloved spouse Constandinos and son Antreas for believing in me when I didn't. You are my rock!*

**Dedication**

This dissertation is lovingly dedicated to the memory of my father Andreas Tamanas, PhD. Although he was my inspiration to pursue a doctoral degree in bioinformatics, he was unable to watch me materialize his wishes. This is for him.

## Table of Contents

## Table of Figures

## Table of Tables

# Abbreviations

| | |
|---|---|
| ARI | Adjusted Rand Index |
| CBR | Compositionally Biased Region |
| CBS | Composition Based Statistics |
| CDS | Coding DNA Sequence |
| DSSP | Dictionary of protein Secondary Structure Pattern |
| ER | Endoplasmic Reticulum |
| ESTs | Expressed Sequence Tags |
| GPI | glycosylphosphatidylinositol |
| HGT | Horizontal Gene Transfer |
| HRP | Homopolymeric Repeat Region |
| IIO | Isolation Index of Organisms |
| LCR | Low Complexity Region |
| MPF | Major Plasmodium Family |
| MSA | Multiple Sequence Alignment |
| MSP | Merozoite Surface Protein |
| ND | Neighborhood Distribution |
| OST | Oligosaccharyltransferase |
| pHMM | profile Hidden Markov Model |
| PDB | Protein Data Bank |
| PfEMP1 | *Plasmodium falciparum* Erythrocyte Membrane Protein 1 |
| RASA | Relative Accessible Surface Area |
| RBC | Red Blood Cell |
| RE | Regular Expression |
| SNCI | Syntenic Neighborhood Conservation Index |
| TRG | Taxonomically Restricted Genes |

## List of associated publications

1. **Tamana S.** and Promponas V. J., 2018, "**An updated view of the oligosaccharyltransferase complex in Plasmodium**", Submitted in Glycobiology
2. **Tamana S.** and Promponas V. J., 2018, "**Dissecting sequence and structural features of compositionally biased regions in the Protein Data Bank***"*, In preparation
3. **Tamana S.** and Promponas V. J., 2018, "**Standard Operating Procedure for computing compositionally biased pangenomes***"*, In preparation
4. **Tamana S.** and Promponas V. J., 2018, "**Unique genes in malaria parasites: a pan-genomic approach**", In preparation
5. Mier P., Paladin L., **Tamana S.,** Petrosian S., Hajdu-Soltész B., Urbanek A., Gruca A., Plewczynski D., Grynberg M., Bernadó P., Gáspári Z., Ouzounis C., Promponas V.J., Kajava A.V., Hancock J.M., Tosatto S., Dosztanyi Z., Andrade-Navarro M.A., "**Disentangling the complexity of low complexity proteins**", Briefings in Bioinformatics (In press).
6. Panayidou S., Georgiade K., Christofi T., **Tamana S.**, Apidianakis G., Promponas V. J., 2018, "**Pseudomonas aeruginosa metabolism as a sizable pool of virulence genes**", in preparation
7. **Tamana S.**, Kirmitzoglou I., and Promponas V.J. "**Sequence features of Compositionally Biased Regions in Three Dimensional Structures**". Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference, 11-13/11/2012

# Chapter 1 – Introduction

## 1.1. Compositionally Biased Regions

*Compositionally Biased Regions (CBRs)* or as also known as *Low Complexity Regions* (LCRs) are found in abundance in nature and in protein databases (Golding, 1999; Wootton and Federhen, 1993). Statistical analyses of amino acid sequence databases globally revealed that approximately 20% and 8% of all known sequences of eukaryotes and non-eukaryotes, respectively, contain at least one cluster of "unusual" amino acid composition (Peng et al., 2015). Until recently, these regions were treated as junk peptides mainly because they tend to conform into non-globular structures (Dunker et al., 2001) - presenting further challenge for trivial experimental procedures to determine their three-dimensional (3D) structures (Altschul et al., 1990; Bannen et al., 2007; Dunker et al., 2001; Romero et al., 2000; Wootton and Federhen, 1993) and require special treatment in fundamental steps of comparative genomic analyses. In fact, the initial scientific interest and research effort for the identification (and consequently filtering) of CBRs was mainly because CBR existence is known to create artifacts (i.e. produce biologically irrelevant hits) in sequence-based database search methods (Altschul et al., 1994; Wootton, 1994; Wootton and Federhen, 1993). Thus, presenting further challenge for their interpretation.

The prevailing definition of LCRs is the one of Wotton & Federhen stating that these regions are clusters of high composition of certain amino acid residue type(s) (Wootton and Federhen, 1993). According to Wootton and Federhen, LCRs are highly non – random regions in protein structures (Wootton, 1994; Wootton and Federhen, 1993) where, many natural macromolecules contain elongated non – globular structures that are difficult to solve with trivial experimental procedures (e.g. X- Ray crystallography, NMR spectroscopy, mutagenesis and other methods) (Wootton, 1994; Wootton and Federhen, 1993). These clusters contain regions of a certain amino acid (e.g. only glycine or serine) or are in a mosaic sequence arrangement, some of which contain regular or irregular short – period tandem repeats (e.g. ERERER or EEKNEKNDEE; (Wootton and Federhen, 1993)). An example highlighting such cases is a protein sequence from the most lethal human malaria parasite, *Plasmodium falciparum* 3D7, where there are marked spans along the protein sequence with composition favoring Glutamic acid, Lysine, Aspartic acid and Asparagine residues (**Figure 1**).

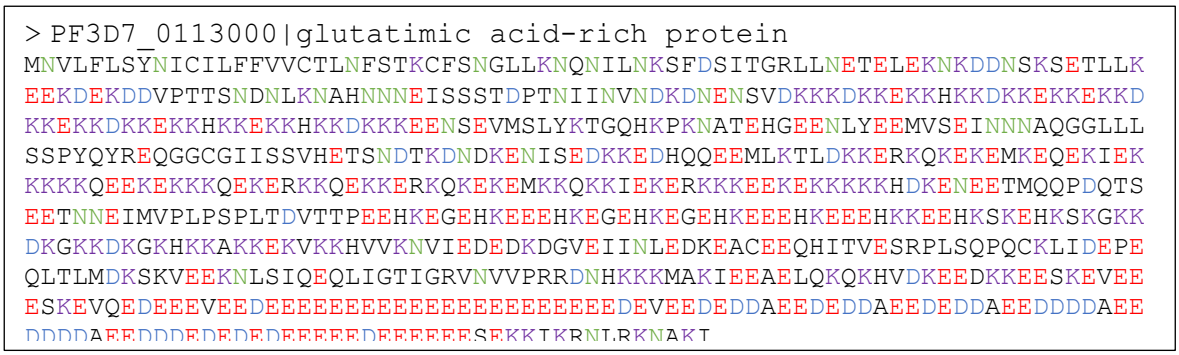


**Figure 1:** An example of an E-rich, K-rich, D-rich and N-rich protein sequence from *Plasmodium falciparum* 3D7 annotated as glutamic acid-rich protein.

Besides, these regions tend to occur as short subsequences of 15 – 50 residues long and what is most surprising, is that they do not resemble the functionally well understood structural proteins such as keratins, collagens and elastins (Saqi, 1995; Wootton, 1994; Wootton and Federhen, 1993). Numerous studies illustrated their tendency to occur in proteins which are important to human disease, for example the arginine–rich region in human immunodeficiency virus rev and tat proteins (Auer et al., 1994), Huntington's disease (Walker, 2007), Parkinson's disease (Gundersen, 2010), Alzheimer's disease (Skrabana et al., 2006) and others. There are also cases when CBRs exist in signaling proteins and assist protein-protein interactions or even act as linkers of different domains within a protein (Dunker et al., 2001; Golding, 1999; Kay et al., 2000; Michelitsch and Weissman, 2000)(Dunker et al., 2001b; Golding, 1999; Kay et al., 2000; Michelitsch and Weissman, 2000).

Moving a step further, Kuznetsov & Hwang (2006), divided the general term of compositional bias into global and local compositional bias. The term of global bias is referring to the entire protein sequence that is consisting of a large amount of specific residue types and becomes one large compositionally biased segment (Kuznetsov and Hwang, 2006). On the other side, the local bias is more about protein sequences that conforms the random independence model of small clusters or upper – represented residue types and that such sites are good candidates for functionally and/or structurally important studies (Kuznetsov and Hwang, 2006).

Another popular definition of proteins containing LCRs is the Intrinsic Disordered proteins. This definition considers segments or even the whole protein that fails to self - fold into a fixed 3D – structure, sometimes even can occur in the native state (Dunker et al., 2001; Romero et al., 2000; Schnell et al., 2007). An extreme example of intrinsically disordered proteins are the two human proteins histone H1 and its nuclear chaperone

prothymosin-α (Pro-Tα) which associate with extremely high affinity complex, but fully retain their structural disorder, long-range flexibility and highly dynamic character (Borgia et al., 2018). However, not all proteins show this lack of defined structure, but they may have local regions of disorder. Take for instance the recently-solved protein structures of the yeast and mammalian oligosaccharyletransferase (OST) complex — a protein complex integrated in the endoplasmic reticulum (ER) membrane and key enzyme of Asparagine-linked glycosylation — organized in subcomplexes with well-defined structures but with local segments of disorder (Bai et al., 2018; Braunger et al., 2018; Wild et al., 2018). Other examples of well-structured CBRs can be found in the human transmembrane protein occludin (PDB id: 1WPA; (Li et al., 2005)), the Glycine-rich antifreeze protein (2PNE; (Pentelute et al., 2008)), and the human RNA binding protein FUS (4FDD; (Zhang and Chook, 2012)). As a consequence, the low complexity regions are local regions of disorder leading to an overlapping definition (Michelitsch and Weissman, 2000; Romero et al., 2000). It is worth mentioning that, even though LCRs often correspond to IDRs, this is not always the case; this issue is discussed in detail in a recent review (Mier et al., 2018).

CBRs are also known as Simple Sequences. However, Simple sequences are considered as subset of CBRs (Albà et al., 2002; Sim and Creamer, 2004) . Huntley & Golding (2002) describe this subset as the perfect repeats of a single amino acid and that there are in excess in eukaryotes but not in prokaryotes (Huntley and Golding, 2002). Moreover, simple sequences are primarily composed of specific amino acids such as Glutamine, Asparagine, Serine, Threonine, Proline, Histidine, Glycine, Alanine, Aspartic acid and Glutamic acid, and also their length is not more that 20 residues long (Huntley and Golding, 2002). A later study, discover that not only some of the simple sequences were conserved but also, were consistent with local structure and function (Sim and Creamer, 2004). Even though the percentage of the conserved simple sequences does not exceed the 11%, these regions showed to obtain functions from substrate binding to structural integrity (Sim and Creamer, 2004). A recent study comparing homo-amino acid repeats of 22 proteomes suggested that, shorter repeats are conserved among orthologs while, proteins with longer repeats (>15 amino acids) found to be unique to the respective organism (Kumar et al., 2016). Lysine-rich CBRs seem to be well conserved among orthologs both in respect to their length and number of occurrences in a protein while, other amino acids such as Glutamic acid, Proline, Serine and Alanine CBRs are generally conserved between orthologs with varying lengths (Kumar et al., 2016). Aspartic acid and

Glutamic acid repeats (i.e. D-/E-rich) have important biological roles owing to their negative charges and underlying properties to interact with metal ions (Chou and Wang, 2015). For instance, Glutamic acid rich proteins serve as markers for the diagnosis of malaria (Kattenberg et al., 2012) and babesia (Mousa et al., 2013) while, Aspartic acid rich proteins are major components of the soluble organic matrix of mollusk shells (Gotliv et al., 2005; Nudelman et al., 2006; Weiner, 1979). Alanine-rich (A-rich) tracts were found to influence the subnuclear targeting properties of the RBM4 CAD in cultured human cells (Chang et al., 2014) and Glycine-Alanine (GA) dipeptide repeats contributes to toxicity in vivo (Ohki et al., 2017).

Furthermore, a recently published review article focus on the different definitions of CBRs as an effort to illustrate the overlaps between the different properties related to CBRs and draw attention to the need of complete annotation of sequences in the databases (Mier et al., 2018). Ultimately, to assist our efforts of gaining a better understanding of the evolution and function of LCRs.

**Structural characteristics of CBRs**

Discovering the tertiary structure of a protein, or the quaternary structure of its complexes, can provide important clues about how a protein performs its function (Lodish et al., 2000).

The most common experimental methods of structure determination include *X-ray crystallography* and *NMR spectroscopy*, both of which can produce information at atomic resolution. X-ray structural models represent the protein structure in a frozen state (i.e. in the ordered crystal lattice). On the other hand, NMR spectroscopy models represent the dynamic protein structure in solution. Also, there is a limitation of the protein's molecular size with NMR spectroscopy which is limited to significantly small macromolecular complexes (<20 kDa) whereas X-ray crystallography can be applied to larger macromolecular complexes (> 100 kDa). Thus, NMR spectroscopy and X-ray crystallography are complementary methods that can provide a wide range of proteins structural properties (Brunger, 1997). Importantly, solved protein structures are usually deposited in the *Protein Data Bank* (PDB, (Berman et al., 2000)), which is a freely available internet resource from which structural data for thousands of proteins can be obtained.

CBRs were often mistaken as 'junk' peptides due to their tendency to conform into non-globular domains or being in disorder state (Dunker et al., 2001; Romero et al., 2000;

Saqi, 1995; Toll-Riera et al., 2012). Such tendency is further accompanied by difficulty in solving the 3D protein structure with commonly-used experimental procedures such as X-Ray crystallography and NMR spectroscopy (Coletta et al., 2010; Crick et al., 2006; Kumar et al., 2017; Romero et al., 2000).

An interesting study by Simon and Hancock (2009) demonstrated that CBRs are predominantly found in unstructured regions and that, approximately 96% of the tracts are enriched in Serine and Glutamic acid and 67% of the tracts are enriched in Alanine (Simon and Hancock, 2009). This is particularly interesting because, as stated by Jorda and Kajava (2010), the level of repeat perfection correlates with their tendency to be unstructured (Jorda and Kajava, 2010). Although spectrometry and computational works showed that despite the absence of hydrophobic residues in CBRs, poly-glutamine and poly-glycine repeats can form collapsed globular structures in dilute aqueous solutions, restricting the chances to form a globular structure in vivo (Crick et al., 2006). However, there are also sequences that in normal conditions are unstructured, but that do not display a single CBR under specific conditions such as temperature factors which are generally the same as the rest of the protein or when chaperoning (Saqi, 1995; Tompa and Kovacs, 2010).

One of the early studies on the structural characteristic of CBRs showed that these segments are predominantly exposed and are mostly helical or coiled (Saqi, 1995; Tamana et al., 2012). Comparisons between highly flexible/disorder structures with ordered proteins revealed noticeable biases in their amino acid compositions (Dunker et al., 2001; Uversky et al., 2000; Uversky and Dunker, 2010) in a way that are composed of less “order-promoting” residue types (Theillet et al., 2013; Uversky, 2013). The order-promoting residue types include mostly hydrophobic amino acids (such as Leucine, Valine and Asparagine) which are commonly found within the hydrophobic cores of foldable proteins as opposed to disorder-promoting residues which are mostly polar and charged residues (such as Glutamic acid, Serine and Proline) that typically located at the surface of foldable proteins (Theillet et al., 2013; Uversky, 2013). Apparently, hydrophobic CBRs are prone to induce either self-aggregation or/and intermolecular interactions with surrounding proteins when exposed and thus, trigger aggregation (Grignaschi et al., 2018).

Nevertheless, diverse approaches, definitions and criteria have been developed for the identification and consequently filtering of CBRs, as an effort to improve sequence

database searches, assist for further investigation of their structural or functional importance by experimental analysis and provide biologically meaningful results.

**Detection and Masking Algorithms**

Through careful research in literature someone can view a wealth of definitions and algorithms developed specifically for the detection and filtering of CBRs. Among these definitions, the complexity varies almost linearly with segment length or whether their structure is random (disorder or not) (Altschul et al., 1994; Bannen et al., 2007; Vasilis J. Promponas et al., 2000; Wootton and Federhen, 1993). In this section, we will provide a brief overview of the algorithms and tools developed for the detection and filtering of CBRs. Some of these algorithms are still being used extensively (e.g. SEG in BLAST suite of tools or CAST) but others left just for literature (e.g. XNU). In Table 1, there is a list of all the detection and masking software's developed so far along with a brief description of what they do.

The *Statistical Analysis of Protein Sequences (SAPS)* software was developed by Brendel V. et al. (1992) and its primary goal was to develop several statistical methods for the evaluation of a variety of protein sequence properties. Furthermore, the program SAPS was designed as a help tool for further experimental analysis. The advantage of this software was the fact it did only that. It marked only regions needed further investigation for their structural or functional importance by experimental analysis (Brendel et al., 1992).

A year later, XNU (Claverie and States, 1993) was published as one of the very first implementations of detection and masking algorithms. The motivation behind its implementation was the fact that CBRs yield many false positives in protein database searches and similarity alignment tools. The implementation of XNU was supposed to correct the output of these programs as it had the ability to identify the erroneous sequences. A key point of the implementation of XNU was the use of the optimal PAM120 substitution matrix for all the database searches where, in the first step, it compares the protein sequence to itself and outputs the scoring of the local alignment using the PAM12O and then estimates their significance according to the statistical analysis proposed by Karlin and Altschul (Karlin and Altschul, 1990). Following the statistical analysis, it uses a Dot – matrix plot in order to find the best local alignments. By default, the best local alignments appear as off – diagonal segments where the low – complexity

segments will be found very close to the main diagonal. In contrast, the high complexity segments will appear at a much greater distance. Furthermore, the significant low complexity regions in the sequences are mapped by projecting the off – diagonal elements (both horizontally and vertically) onto the main diagonal. The final step of XNU algorithm is fixing the low periodicity threshold (by default is ten) in order to operate only on the most spurious low complexity regions (Claverie and States, 1993).

SEG has been the golden standard for identifying and masking low-complexity regions since mid-1990's when it was incorporated into the BLAST suite of programs (Altschul et al., 1990) as a default option. Its sensitivity along with high speed facilitates the elimination of a number of spurious hits produced by BLASTP, practically enhancing its specificity. It is based on a mixture of statistics and probabilities by taking advantage of the Shannon entropy and its statistical properties and can detect most of the non-random regions in protein sequences including homopolymers, short period tandem repeats and aperiodic mosaics of few residue types. SEG is a two stage window based algorithm that accepts 3 user-defined parameters: the initial window length L and two complexity thresholds; the trigger $K_1$ and the extension complexity $K_2$ (Wootton and Federhen, 1993). While SEG performs notably well in most of the cases, its usage as a default option prior to BLAST searching has been the source of skepticism by various authors (Kreil and Ouzounis, 2003; Kuznetsov and Hwang, 2006; Promponas et al., 2000; Wan et al., 2003, p. 200). Most of them agree that SEG masks significantly more residues than those really needed to eliminate spurious hits and in some cases it even masks functionally important regions that would lead to the detection of true homologues (Koonin and Mushegian, 1996). This can be mainly attributed to the origins of SEG as a tool to analyze the local complexity of proteins sequences and to automatically identify non-globular protein domains (Wootton, 1994) both being applications that favor sensitivity of detection upon specificity. Another target of criticism was the usage of equal prior probabilities of residues types in combination with a segmentation threshold derived by random sequences. While this decision is justified by the composition of low-complexity regions which is very different from the general composition of protein databases (Wootton and Federhen, 1993), it also has some important consequences, most notably the independence of the detection from the sequence attributes derived from its amino acid types composition (Wan et al., 2003) and the inability to estimate statistical significances for the detected LCRs (Kuznetsov and Hwang, 2006). Additionally, the usage of sliding

window makes SEG strongly biased in the detection of LCRs with lengths similar to the window length and thus, causes qualitative differences of the biased regions detected (Kreil and Ouzounis, 2003).

An extension to SEG was proposed by Wan and Wootton (2003) who defined the *DSR* measure of reciprocal complexity (Wan et al., 2003). *DSR* takes into account the sequence length as well as the amino acid composition of the database by incorporating scoring schemes (which are actually substitution matrices like BLOSUM62) (Wan et al., 2003).

Another method based on repeat detection is *CARD* (Shin and Kim, 2005). Utilizing the suffix tree data structure and a number of internal parameters *CARD* was proposed as an alternative method for CBR detection (Shin and Kim, 2005).

*Bias* is an extension of *SAPS* approach for the identification of CBRs (Kuznetsov and Hwang, 2006). It can detect CBRs being composed of user-defined sets of residue types using discrete scan statistics and probabilities (Kuznetsov and Hwang, 2006).

A different approach was taken in the development of the *SIMPLE* algorithm by Tautz D. and colleagues (Tautz et al., 1986) and its later extension from Hancock & Armstrong (Hancock and Armstrong, 1994). SIMPLE was designed to measure the cryptic simplicity in DNA sequences in order to detect clustering of tri-nucleotides and tetra-nucleotides above random noise (Albà et al., 2002) .

An iterative dynamic programming approach (*CAST*; (Promponas et al., 2000)) was proposed based on the intuitive idea that CBRs should score exceptionally high when compared to degenerate sequences of homo-amino acid repeats. The main idea of CAST, is to selectively detect CBRs by multiple – pass of the Smith & Waterman algorithm of the query sequence against twenty homopolymers with infinite gap penalties (Promponas et al., 2000). The output of CAST is not only the masked query sequence (for further analysis) but also, the CBRs. The formulation of the problem behind CAST, are that the homopolymeric peptides can be used for the detection of CBRs because by definition, do not contain any actual biological sequence information but can be completely characterized by its monomer type and its length (Promponas et al., 2000). In contrast to other detection and masking algorithms, CAST has the ability to identify the type of residue causing the bias (Promponas et al., 2000). Also, it uses that information in a way that masking can be done in a more subtle and specific manner. The only residue type being masked (i.e. replaced by undefined residue type X) is the over – represented (i.e. has a positive score above a default threshold) in the biased region and all other residues

remain untouched. Numerous studies illustrated the superiority of CAST when applied as a filter prior to BLAST (Altschul et al., 1990) searches due to its selective detection (and masking) properties (Kreil and Ouzounis, 2003; Promponas et al., 2000; Tamana et al., 2012).

A regular expression-based method (*0j.py*; (Wise, 2001)) was also developed for the detection of CBRs. The idea behind *0j.py* is that proteins with CBRs when encoded by regular expressions become compressed.

**Table 1:** A list of algorithms and tools developed for the identification and filtering of CBRs.

| A/A | Algorithm Name | Tool Access | Web link | Short Description | Reference |
|---|---|---|---|---|---|
| **1** | **SAPS** | Free<br>Web-resource | http://brendelgroup.org/bioinformatics2go/SAPS-SSPA.php<br>https://www.ebi.ac.uk/Tools/seqstats/ | Describe several protein sequence statistics for the evaluation of distinctive characteristics of residue content and arrangement in primary structures | (Brendel et al., 1992) |
| **2** | **XNU** | Free | https://github.com/RobertBakaric/XnuFilt | Use of PAM120 scoring matrix for the calculation of complexity | (Claverie and States, 1993) |
| **3** | **SEG** | Free | ftp://ftp.ncbi.nih.gov/pub/seg/seg/ | A 2-pass window-based algorithm where: 1. identifies the LCR, and 2. performs local optimization by masking with X the LCRs | (Wootton and Federhen, 1993) |
| **4** | **CAST** | On request (source code)<br>Web-server | vprobon@ucy.ac.cy<br>http://athina.biol.uoa.gr/CAST/ | Identifies CBRs using dynamic programming. | (Promponas et al., 2000) |
| **5** | **Oj.py** | On request | M.Wise@ccsr.cam.ac.uk | Is a tool for demarcating low complexity protein domains | (Wise, 2001) |
| **6** | **SIMPLE** | Older version<br>Latest version | http://www.biochem.ucl.ac.uk/bsm/SIMPLE<br>https://github.com/john-hancock/SIMPLE-V6 | Facilitates the quantification of the amount of simple sequence in proteins and determines the type of short motifs that show clustering above a certain threshold | (Albà et al., 2002) |
| **7** | **DSR** | On request | hw@ncgr.org | Calculates complexity using reciprocal complexity | (Wan et al., 2003) |
| **8** | **ScanCom** | On request | *Paper does not provide contact email for request. | Calculates the compositional complexity using the linguistic complexity measure | (Nandi et al., 2003) |
| **9** | **CARD** | Free<br>Contact | *This link is not working: http://bioinfo.knu.ac.kr/research/CARD/<br>swshin@bioinfo.knu.ac.kr | Is based on the complexity analysis of subsequences delimited by pair of identical, repeating subsequences | (Shin and Kim, 2005) |
| **10** | **BIAS** | Free | http://lcg.rit.albany.edu/bias/ | Use the discrete scan statistics that provides a highly accurate correction of multiple test to compute analytical estimates of | (Kuznetsov and Hwang, 2006) |

| A/A | Algorithm Name | Tool Access | Web link | Short Description | Reference |
|---|---|---|---|---|---|
| | | | | the significance of each compositionally biased segment | |
| **11** | **GBA** | On request | xli@cise.ufl.edu, tamer@cise.ufl.edu | Is a graph – based algorithm that constructs a graph of the sequence | (Li and Kahveci, 2006) |
| **12** | **SubSeqer** | Free<br>Web-server | *This link is not working: http://compsybio.org/subseqer/<br>http://compsysbio.org/subseqer/ | Is a graph-based approach for the detection and identification of repetitive elements in low – complexity sequences | (He and Parkinson, 2008) |
| **13** | **ANNIE** | Free | http://annie.bii.a-star.edu.sg | Create an automation of the sequence analytic process | (Ooi et al., 2009) |
| **14** | **LPS-annotate** | Free | *This link is not working: http://libaio.biol.mcgill.ca/lps-annotate.html | The algorithm defines compositional bias through a thorough search for lowest-probability subsequences (LPSs; Low Probability Sequences) and serves as workbench of tools now available to molecular biologists to generate hypotheses and inferences about the proteins that they are investigating. | (Harbi et al., 2011) |
| **15** | **LCReXXXplorer** | Free | http://repeat.biol.ucy.ac.cy/fgb2/gbrowse/swissprot/ | Is a web platform to search, visualize and share data for low complexity regions in protein sequences. LCR-eXXXplorer offers tools for displaying LCRs from the UniProt/SwissProt knowledgebase, in combination with other relevant protein features, predicted or experimentally verified. Also, users may perform queries against a custom designed sequence/LCR-centric database. | (Kirmitzoglou and Promponas, 2015) |
| **16** | **fLPS** | On-line resource<br>Github download | http://biology.mcgill.ca/faculty/harrison/flps.html<br>https://github.com/pmharrison/flps | fLPS can readily handle very large protein data sets, e.g. stemming from metagenomic projects. It is useful in searching for proteins with similar CBRs, and for making functional inferences about CBRs for a protein of interest. | (Harrison, 2017) |

A different approach based on the so-called, linguistic complexity of sequences is illustrated in the software named *ScanCom* (Nandi et al., 2003). In their method (ScanCom), dimer "word" counts were used for identifying segments with extreme composition.

ANNIE is one of the first tools that automates an essential part of the sequence analysis process and integrates over twenty function prediction algorithms (including CBRs detection algorithms) proven to be reliable and indispensable for protein sequence annotation. The results of individual algorithms can be accessed separately or displayed together through a web-sequence viewer and at the same time, users can assess certain features of a set of sequences or their taxonomic distribution (Ooi et al., 2009).

Later studies (He and Parkinson, 2008; Li and Kahveci, 2006) taking advantage of novel data structures for representing sequence data, created software based on graphs for the identification of CBRs. The *GBA* algorithm rejects the Shannon entropy as a good complexity measure for protein sequences and relies on a new complexity measure that represents CBRs in a graph (Li and Kahveci, 2006). Another method for the identification and characterization of CBRs using graphical visualization methods is *SubSeqer*. The main idea that inspired the authors to the development of *SubSeqer* were given through protein-protein interaction studies (He and Parkinson, 2008; Li and Kahveci, 2006).

More recent additions to the workbench of CBR detection and filtering tools are LPS-annotate (Harbi et al., 2011), LCR-eXXXplorer (Kirmitzoglou and Promponas, 2015) and fLPS (Harrison, 2017). The database server LPS-annotate was developed specifically for the analysis and annotation of CBRs and protein disorder in protein sequences (Harbi et al., 2011). Briefly, the LPS-annotate algorithm defines compositional bias through a search for lowest-probability subsequences (LPSs) where, users can annotate CBRs in protein or nucleotide sequences and search a database of pre-calculated protein-CBRs for functional hypotheses and their protein disorder propensities (Harbi et al., 2011). LCR-eXXXplorer is a web-platform to search, visualize and share data for LCRs in protein sequences. It offers tools for displaying pre-computed LCRs from the UniProt/SwissProt knowledgebase (The UniProt Consortium, 2017), in combination with other relevant protein features, predicted or experimentally verified (Kirmitzoglou and Promponas, 2015). Additionally, users may perform powerful queries against a custom designed sequence/LCR-centric database and can download the complete set of masked sequences in FASTA formatted files, the complete set of annotations in GFF3 format or a CSV formatted table with LCR

statistics for each sequence in the database (Kirmitzoglou and Promponas, 2015). Moving to the metagenomic era, the latest addition of CBR-detection fLPs algorithm, can handle very large protein data sets, such as the whole UniProt/TrEMBL (The UniProt Consortium, 2017), in approximately one hour (Harrison, 2017). It discovers both single-residue and multiple-residue biases through a process of probability minimization and window-based searches where efficiency measures are taken to avoid or delay probability calculations (Harrison, 2017).

Regardless which definition of CBRs we use, it is evident that all these algorithms were developed within a specific content, the detection and masking of regions with an unusual amino acid composition in single species genomes. However, with the emergence of high-throughput 'next-generation' sequencing technologies, an incredible source of already complete sequenced genomes and with the number of newly sequenced genomes steadily increasing, it makes sense to re-think the idea of how CBRs affecting the computations of heavily biased genomes into a single pan-genome, such as the malaria parasites.

**Pan-genome analyses and Comparative genomics**

Comparative genomics are defined as "a large-scale holistic approach that compares two or more genomes to discover the similarities and differences between the genomes, and to study the biology of the individual genomes" (Wei et al., 2002). Over the past two decades, whole-genome analyses and comparative genomics revealed an astonishing level of genomic variation across the tree of life and thus, challenging the value of studying single reference genomes (Offord, 2016). Hence, highlighting the significance of pan-genomes in studying the evolutionary forces that shape modern genes, especially in our effort to develop universal vaccines that could provide protection against all strains in a species or even against several related species (Offord, 2016).

A pan-genome is defined as the union of all genes encoded in all the strains in a particular clade (Tettelin et al., 2005). The original concept of pan-genome was conceived by Sigaux while he was trying to define new tests that will be useful for diagnostic procedures in clinical laboratories and new targets for biological treatments of tumors (Sigaux, 2000). Few years later, Tettelin and colleagues defined *Streptococcus agalactiae* pan-genome as the combination of a 'core' genome, containing genes present in all strains (accounting for approximately 80% of any single genome), and a 'dispensable' genome (also known as

flexible or accessory genome) composed of genes absent from one or more of the strains (Tettelin et al., 2005). In their study, using mathematical extrapolation they suggested that the gene reservoir of the *S. agalactiae* pan-genome is huge and that new unique genes will continue to be identified even after sequencing hundreds of genomes. However, a study using 44 sequenced strains of *Streptococcus pneumoniae*, observed that fewer new genes were discovered with each new genome sequenced and that the predicted number of new genes dropped to zero when the number of genomes exceeds 50 (Donati et al., 2010). In fact, the main source of new genes in *S. pneumoniae* was *Streptococcus mitis* from which genes were horizontally transferred (Donati et al., 2010). Thus, suggesting that there are two generalized types of pan-genomes, the "closed pan-genome" and the "open pan-genome", categorized by the number of new genes added to the pan-genome per sequenced genome (Vernikos et al., 2015). Species with a closed pan-genome would have very few genes added per sequenced genome even after sequencing many strains, while species with an open pan-genome could have enough genes added per additional sequenced genome (Vernikos et al., 2015).

To date, over 50 bacterial species have been re-analyzed using the pan-genome framework, including *Escherichia coli* and *Chlamydiales* (Chaudhari et al., 2016; F. E. Psomopoulos et al., 2012). Conversely, pan-genome studies in malaria parasites are limited to few *Plasmodium* species (Ansari et al., 2016; Bensch et al., 2016; Borner et al., 2016; Cai et al., 2010; Chaudhry et al., 2018; Rutledge et al., 2017) mainly due to their extremely biased genomes and the lack of an optimal comparative genomics pipeline assessing the effects of CBRs in computing the *Plasmodium* pan-genome. Specifically, the initial comparative genomic analysis in *Plasmodium* was initiated by gene mapping studies on separated chromosomes (Carlton et al., 2002), to infer the parasite's evolution and age (Escalante and Ayala, 1994; Hall et al., 2005; Martinsen et al., 2008), to determine protein motifs and function (Florent et al., 2010; Tyagi et al., 2011), to identify orthologues and generate synteny maps (Cai et al., 2012, 2010) and importantly, to identify candidate drug/vaccine targets (Birkholtz et al., 2008; Ludin et al., 2012). Chaudhry S. and colleagues performed comparative genomics as an effort to understand the evolutionary and functional significance of CBRs in *Plasmodium* genomes, whether driven by natural selection or neutral evolution (Chaudhry et al., 2018).

**Unique proteins**

The classical model of evolution proposed that species divergence, adaptation or extinction results from stepwise genetic changes (such as mutations, damage or replication errors in an organisms' DNA) that ultimately shaped early lifeforms into more complex biological systems (Tautz and Domazet-Lošo, 2011) . As this genetic variation of a population drifts randomly over generations, natural selection gradually leads traits to become more or less common based on the relative reproductive success of organisms with those traits (Kaessmann, 2010).

An interesting subset of genes contributing both in genome and phenotypic evolution of all known organisms, are the so-called *Orphan* or *De novo* genes which are *unique* to a specific taxonomic level (usually restricted to a specific species strain, e.g. *P. falciparum* 3D7) (Khalturin et al., 2009). These unique genes differ from the other genes of organisms as they do not have any detectable homologues in other lineages and in most cases, do not follow the classical model of evolution. The Orphan genes, at first, it was speculated that originated from gene duplication, but several studies indicated that may arise through a variety of mechanisms, such as horizontal gene transfer, duplication and rapid divergence, or de novo origination (Cai et al., 2008; Donoghue et al., 2011; Kaessmann, 2010; Reinhardt et al., 2013; Tautz and Domazet-Lošo, 2011; Zhou et al., 2008) .

The term Orphan genes was first used in 1996, when the yeast whole-genome sequencing project began and accounted, at the time, approximately 26% of the yeast genome (Tautz and Domazet-Lošo, 2011) . However, due to the small number of fully sequenced genomes at the time, the origin of orphan genes was thought as lack of sequencing data rather than due to true lack of homology (Khalturin et al., 2009). However, as more genomes were fully sequenced and contributed in the study of evolution of species, the percentage of orphan genes was persistent to approximately 10%-30% of genomes suggesting that either, these are fast evolving genes leaving few traces back to a common ancestor or originated from non-coding genomic sequences (Martinsen et al., 2008; Outlaw and Ricklefs, 2011; Perkins and Schall, 2002; Schaer et al., 2013). An alternative definition for Orphan as Taxonomically Restricted Genes (TRGs) was given by G. A. Wilson and colleagues as an attempt to capture their strain/species specific origin (Wilson et al., 2005).

Ohno and Epplen (1983) were the first authors suggesting that CBRs play a crucial role in the generation of genetic variation (Ohno and Epplen, 1983). Specifically, they speculated

that the first protein coding sequences were highly repetitive. The expansion of repetitive tracts along with the accumulation of base substitutions, insertions and deletions would eventually diversify the initially repetitive sequences leaving few traces back to the original sequence (Ohno and Epplen, 1983). Recently evolved genes, seem to present a higher fraction of CBRs in their coding sequence than older genes (Albà and Castresana, 2005; Toll-Riera et al., 2012) . Moreover, young genes evolve faster and experience more variable selection pressures than older ones thus, this higher fraction of CBRs in proteins encoded by these genes suggests that DNA slippage induced changes may be better tolerated in this type of proteins (Nishizawa et al., 1999; Vishnoi et al., 2010). These observations seem to support the idea that, the evolutionary model of protein coding genes is more likely to be a dynamic model in which, first, they acquire new regions by repeat expansion and, secondly, they build-up substitutions within these repetitive sequences (Toll-Riera et al., 2012). Thus, after millions of years of evolution what we actually see is the degenerated form of the sequence due to the build-up mutations.

Remarkably, the study of evolution and shaping of unique genes accelerated after a study in 2008, found a yeast protein with known function, to have evolved de novo from non-coding sequences whose homology was still detectable in sister species (Cai et al., 2008). Then, a study in 2009, where an orphan gene was found to regulate an internal biological network in *Arabidopsis thaliana* (Li et al., 2009), leading to a genome-wide study of the extent and evolutionary origins of unique genes in plants in 2011 (Donoghue et al., 2011). An early *Plasmodium* whole genome study suggested that chromosomal rearrangements in the core regions are involved in the generation and subsequent dispersal of *P. falciparum*-specific gene families (Kooij et al., 2005). Studies of species-specific genes, such as the RIFIN or PfEMP1 gene families of *P. falciparum*, revealed that multi-gene duplications followed by neo-functionalization could provide a mechanism for malaria parasites to evolve lineage-specific surface antigens with tissue-specific activities (Cai et al., 2012). Transcriptomic evidence demonstrated that a group of species-specific proteins expressed throughout the 48h intraerythrocytic cycle of *P. vivax* are essential regarding virulence and host pathogen interactions (Bozdech et al., 2008; Zhu et al., 2016).

It has been argued, that unique genes could be spurious non-functional Open Reading Frames (ORFs) mainly because, these genes tend to be very short (~6 times shorter than mature genes), simpler in codon usage and amino acid composition (Arendsee et al., 2014) and that mostly encode intrinsically disordered proteins (Mukherjee et al., 2015;

Verster et al., 2017; Wilson et al., 2017). As this contradicts the studies proving the biological significance of Orphan genes, computational methods and statistics were developed in an effort to distinguish “spurious” Orphan genes from authentic Orphans (Cuadrat et al., 2014; Fukuchi and Nishikawa, 2004; Wilson et al., 2007; Zheng et al., 2005).

Another possible origin of unique genes could be through accidental contamination (i.e. introduction of “foreign” material) of the species genome (Salter et al., 2014; Weiss et al., 2014). A contaminant sequence is defined as one that does not actually represent the genetic information from the biological source organism/organelle because it contains one or more sequence segments of foreign origin. While huge technological advances improved high-throughput sequencing sensitivity, it is still possible the introduction of contaminating DNA during sample preparation (Salter et al., 2014). The most common sources of contamination are cloning vectors (e.g.: plasmid, phage, cosmid, BAC, PAC, YAC) which, unless they are identified and removed, will result in a contaminated sequence. Unintended events can also introduce contamination from other sources such as transposable elements and insertion sequences (National Center of Biotechnology Information, 2018; Weiss et al., 2014).

An intriguing hypothesis is that many human health conditions can be determined more by microbial DNA than human DNA (Lozupone et al., 2013; Weiss et al., 2014). This hypothesis is formed by correlative and experimental studies on human phenotype and mouse models trying to assess how the microbiome affects numerous health conditions such as, obesity and multiple sclerosis (Berer et al., 2011; Vijay-Kumar et al., 2010). Thus, the presence of contaminated sequences could be very beneficial towards our efforts to close important biological gaps (Weiss et al., 2014).

Nevertheless, if samples are not collected, processed or analyzed correctly this could lead to time and effort wasted on meaningless analyses, mis-assembly of sequence contigs and false clustering of Expressed Sequence Tags (ESTs), sequences contaminated with the same foreign sequence can be aligned via the shared foreign segment and finally, pollution of public databases (National Center of Biotechnology Information, 2018; Weiss et al., 2014).

## 1.2. Malaria parasites

Malaria is a mosquito-borne infectious disease of humans and other animals caused by parasitic protozoans of the genus *Plasmodium* of the phylum of Apicomplexans ((Zilversmit and Perkins, 2008); **Figure 2**). Most species of this phylum share a non-photosynthetic plastid organelle, known as *apicoplast* that arose from the endosymbiosis with a red algae (Kissinger and DeBarry, 2011). Scientists believe that the apicoplast may hold the key for the eradication of malaria, as they are sensitive to several anti-bacterials and herbicides (Ralph et al., 2004).

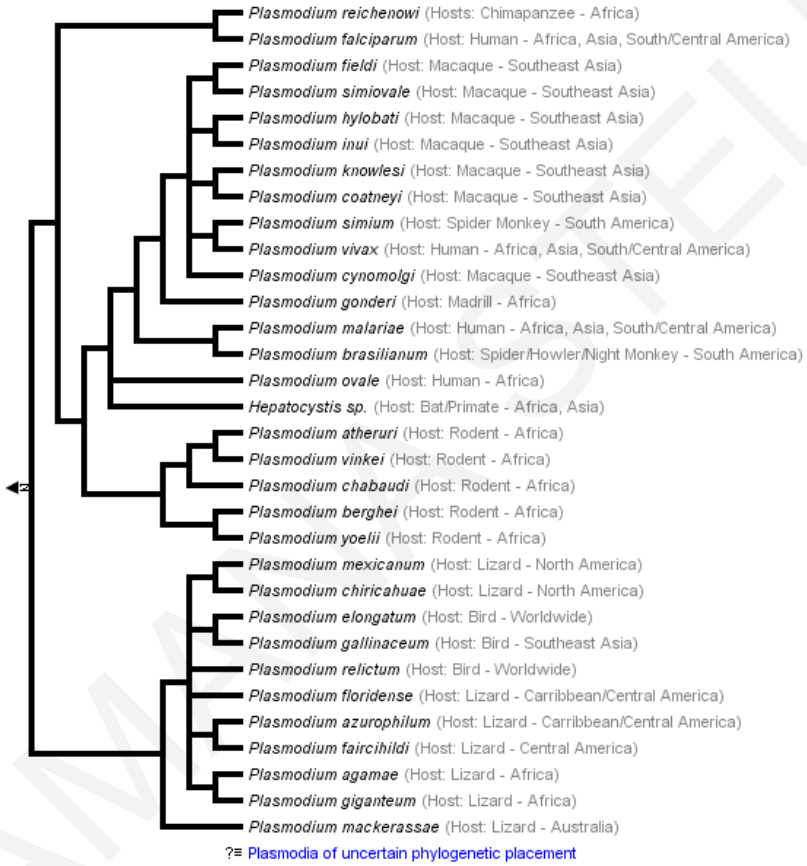



**Figure 2:** Phylogeny of Malaria Parasites (Zilversmit and Perkins, 2008).

The disease is transmitted via a bite from an infected female *Anopheles* mosquito (Figure 3), which introduces the organisms from its saliva into the host's circulatory system (Carlton, 2006). In the blood, the parasites travel to the liver in order to mature and reproduce (Carlton, 2006). The complex life cycle of *Plasmodium* species (**Figure 3**), which alternates between a sexual stage within the mosquito and an asexual stage within the mammalian tissues and blood (Hall et al., 2005), classifies these species as very complex organisms to thoroughly study their biology *in vivo* (Beck, 2006; Hall et al., 2005).

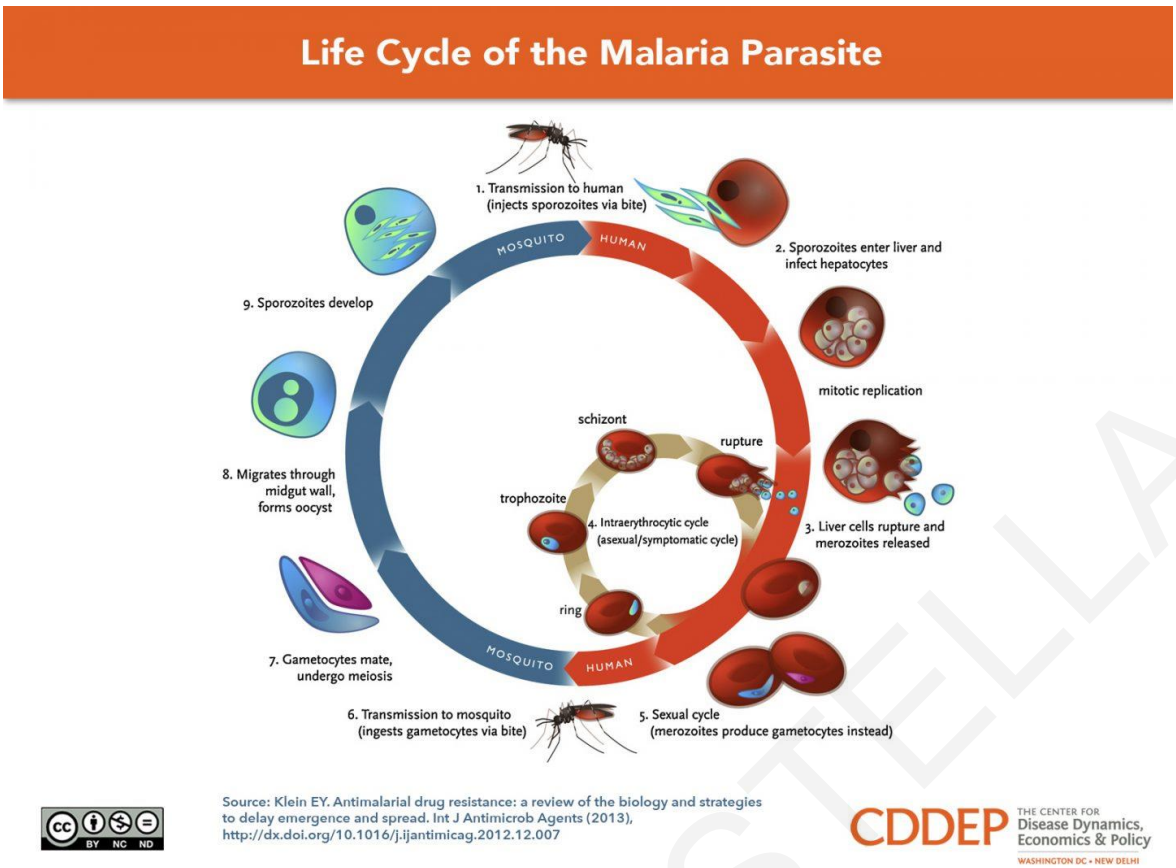



**Figure 3:** An illustration of the life cycle of the malaria parasites. Image source: (Klein, 2013)

Malaria causes symptoms that typically include fever and headache, which in severe cases can progress to coma or death (Fairhurst et al., 2012) and is widespread in tropical and sub-tropical areas of Africa, South America and Asia (World Health Organization, 2017). Human malaria is caused by four species, namely: *Plasmodium falciparum (highest fatality)*, *Plasmodium vivax*, *Plasmodium malariae* and *Plasmodium ovale*. In addition, *Plasmodium knowlesi*, which primarily infects monkeys, was also found to infect humans (Frech and Chen, 2011; Lee et al., 2011). Similarly, other human malaria parasites including *P. vivax*, *P. cynomolgi*, *P. malariae*, *P. simium* and *P. basilium* seem to switch hosts, hence, forewarn for new risks for human public health (Cai et al., 2012; Zilversmit and Perkins, 2008). An interesting issue and something to be seriously concerned, is the fact that the observed rise in monkey-malaria infections among humans comes with the effect of destroying the natural habitat of simians (Law, 2018). Scientists believe that deforestation and close proximity of simians to humans is making it easier for some simian infections to switch between macaques and humans. The reasoning is that it increases a person's chances of getting bitten by mosquitoes infected with malaria (Fornace et al., 2016). If people are constantly at risk of exposure to monkey malaria

parasites, then it complicates our efforts to successfully eradicating malaria (mainly due to the broad reservoir of the parasites within the simian population). Current anti-malaria fighting and treatment approaches revolves around drugs and using bed nets (World Health Organization, 2018) but as these measures cannot be applied in wildlife, all our efforts for successfully eradicating malaria seem foredoom.

Studies performed in a variety of species of this phylum revealed relatively small genome sizes with significantly reduced numbers of protein-encoding gene (Carlton et al., 2002; Malcolm J. Gardner et al., 2002; Kissinger and DeBarry, 2011). *Plasmodium* genome is divided into 14 chromosomes with high Adenine and Thymine (A+T) content (approximately 80% in *P. falciparum*; (Carlton, 2006; Malcolm J. Gardner et al., 2002)). This A+T richness beside the induced difficulty in genome sequencing projects and cloning in heterologous vector systems, results in heavily compositionally biased protein sequences, needing special treatment for the sequence comparison step which is fundamental in comparative genomic analyses (Malcolm J. Gardner et al., 2002; Promponas et al., 2000). Thus, without a thorough understanding of the biology of these parasites we are limiting our options for developing a highly effective vaccine and successfully eradicating malaria.

***Plasmodium falciparum* evolutionary origin**

One intriguing issue is the evolutionary origin of the most lethal human malaria parasite, *P. falciparum*, where two are the most prevalent hypotheses about the most probable descent of the parasite (Hagner et al., 2007). Earlier studies proposed an avian origin as a result of a relatively recent host switch (Escalante et al., 1998; McCutchan et al., 1996; Waters et al., 1991), while more recent studies have found evidence supporting that *P. falciparum* is closely related with the primate-infecting malaria parasites (Borner et al., 2016; Martinsen et al., 2008; Outlaw and Ricklefs, 2011; Perkins and Schall, 2002; Schaer et al., 2013).

The avian-origin hypothesis began by the study of Waters A. P and colleagues (1991), one of the early molecular phylogenies (Escalante et al., 1998; McCutchan et al., 1996; Waters et al., 1993, 1991) but due to insufficient taxon sampling and choice of genes, soon it was rejected as new evidence based on a cytochrome b phylogeny, larger sample size of vertebrate parasites and an outgroup from a sister family placed *P. falciparum* within the clade of mammalian parasites (Perkins and Schall, 2002; Qari et al., 1996). Furthermore,

this analysis using the Shimodaira-Hasegawa test, a phylogenetic tree-based method that calculates the likelihood of alternate trees, tested the avian origin of *P. falciparum* and rejected it (Perkins and Schall, 2002).

However, a recent phylogenetic study, performed by Bensch and colleagues (2016), using the newly sequenced genome of *Haemoproteus tartakovskyi* and novel transcriptome sequences of the avian parasite *P. ashfordi* rejected the mammalian-ancestor hypothesis for the *P. falciparum*, as genes and genomic features shared between *P. falciparum* and avian parasites are absent in other mammalian-infecting malaria parasites (Bensch et al., 2016). Additionally, a study using the core genome of 10 completely sequenced *Plasmodium* species along with a new reference genome of *P. malariae* and a manually curated draft of *P. ovale* genome (the other two human-infecting malaria parasites) placed a split between the *Plasmodium* progenitor and the avian *P. gallicaneum* (Rutledge et al., 2017). The phylogenetic analysis of this recent study suggests that, avian parasite speciation event predated *P. falciparum* speciation and thus, reinforcing the hypothesis that *P. falciparum* shares a common ancestor with the avian parasites (Rutledge et al., 2017).

Up to date, conclusive evidence indicates that the only species closely related to *P. falciparum* is *P. reichenowi* (Escalante and Ayala, 1994; Martinsen et al., 2008; Perkins and Schall, 2002), and the two likely diverged from each other between 5 and 8 million years ago based on fossil dates of the human-chimpanzee split (Escalante and Ayala, 1994).

Nevertheless, as more complete genome sequences of several *Plasmodium* species infecting rodents, simians, humans and avian added in the repertoire, it provides sufficient data for designing optimal strategies for comparative genomic analyses and thus, providing a step forward for a better understanding of the evolutionary history of this elusive species.

**Major *Plasmodium* Protein Families**

A gene family is defined as a group of genes/proteins that share significant sequence similarity and a common evolutionary history (Carlton, 2006). Proteins within paralogous gene families preserve their structure and maintain similar or identical biochemical functions across evolutionary distances (Carlton, 2006). Study of *Plasmodium* Major Protein Families (MPFs) could highlight important features of both their evolutionary history and their medical importance as drugs/vaccine candidates.

**The vir/yir/cir/bir/kir/pir gene family**

This is the largest variant gene family known to date, present in human-, simians- and rodent-infecting malaria parasites predicted to be involved in antigenic variation. It was first described in *P. vivax* (the *vir* family; (del Portillo et al., 2001)) and latterly in *P. yoelii* (the *yir* family; (Carlton et al., 2002)), *P. berghei* (the *bir* family; (Janssen et al., 2002)), *P. chabaudi* (the *cir* family; (Janssen et al., 2002)) and *P. knowlesi* (the *kir* family; (Carlton, 2006)).

The *vir* genes are more divergent than all rodent *yir* genes, sharing 20-30% amino acid identity in predicted polypeptides. Janssen and colleagues (2004) carried out a cross-species phylogenetic analysis using 157 amino acid pir sequences from *P. vivax*, *P. knowlesi*, *P. yoelii*, *P. berghei* and *P. chabaudi* and suggested that this gene super-family may also include the *P. falciparum* multi-gene families *rifin* and *stevor* (Janssen et al., 2004). Based on these findings the authors suggested that the *rifin* and *pir* genes share an ancestral gene sequence, although its unknown whether any functional, regulatory or export mechanisms have been conserved. However, true homologs of this family (initially thought not to exist in *P. falciparum* due the *var* gene family; (Craig and Scherf, 2001; Janssen et al., 2002)), were found almost a decade after their characterization (Frech and Chen, 2013).

The rodent malaria *yir* sequences cluster together in the network, with *bir* and *yir* sequences being closest (interspersed). All three are more distant from *vir* and *pir*, which arise from the same node, reflecting the parasites recent evolutionary divergence from the simian and human infecting species (Hayakawa et al., 2008).

**SICAvar gene family**

SICAvar is a *P. knowlesi*-specific gene family which is expressed on the surface of infected erythrocytes and is implicated in antigenic variation in this species (Pain et al., 2008). No significant sequence similarity exists between the *var* and *SICAvar* genes (Thomas D. Otto et al., 2014; Pain et al., 2008).

**SURFIN gene family**

SURFIN gene family is encoded by ten surf genes, including three pseudogenes and located within or close to the subtelomeres of five *P. falciparum* chromosomes (Winter et al., 2005). *SURFIN*s show structural and sequence similarities with *SICAvar*, *vir* and *var*

genes, and have been implicated in the invasion of erythrocytes by *Plasmodium* during the merozoite phase of its lifecycle (Winter et al., 2005).

### *P. falciparum* Erythrocyte Membrane Protein 1 gene family

Genome comparison of human and non-human malaria parasites reveals species subset-specific genes potentially linked to the human disease (Frech and Chen, 2011). Early comparative genomics analyses of *Plasmodium* genomes showed that genes mediating parasite-host interactions are frequently restricted to a single *Plasmodium* species (species-specific) or restricted to a subset of *Plasmodium* species (TRG).

*P. falciparum erythrocyte membrane protein 1 (PfEMP1)* is one of the best-studied and clinically most important MPF. Different isoforms are encoded by approximately 60 members of the *P. falciparum*-specific *var* gene family (Bull and Abdi, 2016; Malcolm J. Gardner et al., 2002; Su et al., 1995). PfEMP1 proteins are expressed at the surface of infected Red Blood Cells (RBCs) where they mediate adhesion to both uninfected erythrocytes and host endothelial cells (Su et al., 1995) and are the major cause of the severe *P. falciparum* pathology.

### Merozoite Surface Protein 3

*Merozoite Surface Protein 3* (MSP3) is an important member and vaccine candidate of a multigene family expressed during the merozoite blood-stage infection of *Plasmodium* species (Beeson et al., 2016; Oeuvray et al., 1994). Screening of *P. falciparum* genome-wide expression library using Antibody Dependent Cellular Inhibition (ADCI) assay demonstrated that MSP3 is targeted by the naturally occurring host antibodies proved to be lethal for the parasite (Beeson et al., 2016; Imam et al., 2014). Since its characterization in *P. falciparum* strains (McColl et al., 1994; Oeuvray et al., 1994; Pattaradilokrat et al., 2016) studies showed that it is a 48kDa protein with a highly conserved C-terminal domain, possesses a Duffy-binding like domain and that it binds to RBCs, suggesting a functional role of assisting the merozoite attachment to RBCs (Beeson et al., 2016; Sakamoto et al., 2012; G. P. Singh et al., 2004).

Most members of the MSP3 gene family (MSP3.1–MSP3.8) contain the *NLRNA/G* signature N-terminal peptide but it appears that 2 (MSP3.5/MSP3.6) out the 8 MSP3 proteins lack the conserved C-terminal domain and share less than 54% sequence similarity (Singh et al., 2009). Additionally, *P. vivax* and *P. knowlesi* MSP3 orthologues and

paralogues have a heptad Alanine (A)-rich domain coiled-coil domain similar to a shorter coiled-coil region of the *P. falciparum* MSP3.1 protein (McColl et al., 1994).

**OST complex subunits in protists**

Asparagine-linked (N-linked) glycosylation is one of the most common protein post-translational modifications throughout all domains of life (Kelleher and Gilmore, 2006; Schwarz and Aebi, 2011). In eukaryotes, N-linked glycosylation takes place in the endoplasmic reticulum (ER) and oligosaccharyltransferase (OST) ¬– a protein complex integrated in the ER membrane (Li et al., 2008; Pfeffer et al., 2014) – is a key enzyme in this process. OST catalyzes the co-translational transfer of oligosaccharides to the side chains of asparagine residues of transmembrane or exported proteins, provided they are found in 'sequons' determined by the consensus sequence N-X-S/T (where X can be any amino acid residue except proline); serine or threonine occupy the third position, with only some rare exceptions (Aebi, 2013; Chavan and Lennarz, 2006).

The OST complex has been extensively studied in the yeast Saccharomyces cerevisiae. Specifically, in yeast OST appears to function as a hetero-octameric complex composed of the catalytic subunit Stt3p accompanied by Ost1p, Ost2p, Ost4p, Ost5p, Wbp1p, Swp1p and one of Ost3p or Ost6p, all of which are integral membrane proteins of the ER (Kelleher and Gilmore, 2006). Homologous OST subunits have been identified in mammals (Kelleher and Gilmore, 2006), where additional subunits DC2 and keratinocyte-associated protein 2 (KCP2) have been also characterized (Shibatani et al., 2005) and shown to be associated with the STT3A complex (Shrimal et al., 2017). Possible homologs of the yeast OST subunits have been identified across almost all eukaryotes with completely sequenced genomes with only a few exceptions (Kelleher and Gilmore, 2006; Figure 4a); this identification seems to have been conducted mainly by querying protein sequence databases with sequences of known OST subunits.

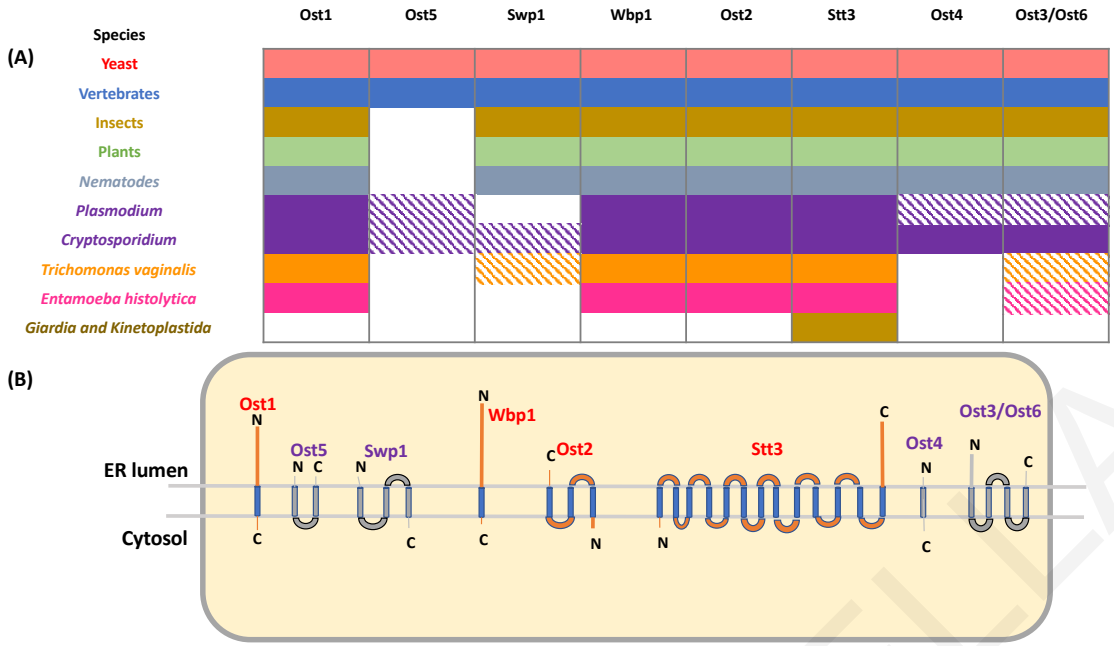



**Figure 4:** Membrane topology and conservation of OST subunits among eukaryotes. (A) The current knowledge on subunit distribution along major eukaryotic groups based on Kelleher and Gilmore (2006) and literature search. The patterned boxes indicate the OST subunits for which evidence is presented in this work. (B) Transmembrane topology of OST subunits based on recently determined 3D structures of the yeast (Wild et al., 2018) and canine (Braunger et al., 2018) OST complex (segment lengths not in scale). The yeast nomenclature for naming OST subunits has been followed. OST subunits already known to be encoded in Plasmodium are shown in red and the remaining in purple.

Recent works describing atomic resolution, cryo–electron microscopy structures of the yeast (Bai et al., 2018; Wild et al., 2018) and canine (Braunger et al., 2018) OST complex have shed light into the transmembrane topology of OST subunits (Figure 4b), their organization into subcomplexes and interactions, enhancing our understanding of the function of OST (Shrimal and Gilmore, 2018).

Interestingly, only four OST subunits (namely Wbp1, Ost2, Stt3 and Ost1) were reported to be encoded in the *Plasmodium falciparum* genome (Kelleher and Gilmore, 2006). This view regarding the composition of the OST complex in malaria parasites has dominated the literature for more than a decade (Aebi, 2013; Breitling and Aebi, 2013; Cova et al., 2015; Lombard, 2016; Shrimal and Gilmore, 2018).

The potential importance of N-linked glycosylation in *Plasmodium* spp. remains a controversial issue (Doerig et al., 2015; Macedo et al., 2010). The absence of some of the OST subunits from plasmodial species is in agreement with previous works providing evidence that most N-linked glycan (Alg) glycosyltransferases –functioning either in the cytosol or in the ER lumen– have no detectable homologs in *Plasmodium* and other protists; this fact has been attributed to their secondary loss from a common ancestor

(Samuelson et al., 2005). A complementary observation is the absence of genes coding for calnexin and calreticulin, which participate in the protein folding quality control of glycoproteins in the ER (Banerjee et al., 2007; Samuelson et al., 2005; von Itzstein et al., 2008). One may argue that 'dispensable' OST subunits may have been lost during the evolution of *Plasmodium* and (possibly) explain the low numbers of N-linked glycosylated proteins (Kimura et al., 1996) in this genus. However, recent biochemical and morphological experiments have confirmed the presence of sugar nucleotides (e.g. UDP-GlcNAc, Sanz et al., 2013) and short N-glycans (GlcNAc or GlcNAc2, (Bushkin et al., 2010; Samuelson and Robbins, 2015)) in different *P. falciparum* life stages.

The reported absence of the aforementioned subunits in malarial parasites has apparent implications in both our understanding of the evolution of the N-linked glycosylation machinery but also on research efforts to elucidate the structure and function of the OST complex in this genus. In this work, we illustrate that three out of the four 'missing' OST subunits (namely Ost3p/Ost6p, Ost4 and Ost5) are actually encoded in completely sequenced *Plasmodium* genomes, based on elaborate sequence database searches, findings which are further corroborated by publicly available microarray, EST and RNAseq transcriptomic data.

### 1.3. Hypothesis and Objectives

Overcoming the early reluctance regarding the biological properties of CBRs, a wealth of definitions and algorithms have already been proposed (Table 1), reminiscent of a "detection babel". Although, this build-up research effort brought to the forefront the biological significance of CBRs (Sreenivas Chavali et al., 2017) at the same time, perplexity on which tool or computational pipeline is best to follow when dealing with heavily biased genomes, such as malaria parasites genomes, is still evident.

Today, with the existing knowledge of the biological significance of CBRs, the availability of several high-quality computational methods and with the number of fully sequenced genomes steadily increasing, we have the opportunity to unravel (i) how CBRs affect the computations of heavily biased pan-genomes, (ii) the role of CBRs in evolutionary behavior of the gene/protein families under study and (iii) CBR's preferences in structural conformations for novel structure-function predictions.

Thus, in this study, we propose an optimal *in silico* strategy focusing on:

**1. Issues in computing the extremely biased pan-genomes**

We investigate the effect of handling CBRs in the computation of the pan-genome, utilizing it with completely sequenced *Plasmodium* species. In this work package, we enlist commonly used algorithms for the detection and masking of CBRs, sequence comparison and clustering.

Here, we follow a pan-genome analysis pipeline (F. E. Psomopoulos et al., 2012) by merging nine complete sequenced *Plasmodium* genomes into a single data file and setting different modes for sequence comparison and clustering.

**2. Plasmodium pan-genome analysis**

Following the computation of the *Plasmodium* pan-genome and the identification of the most reliable protein families, we set to investigate various aspects of the *Plasmodia*. In details, we will focus on key aspects of the pan-genome and explore multiple features of the *Plasmodium* proteome. We anticipate that in this set of experiments we will provide key findings regarding unique/multimember protein families and evolutionary history of the parasite CBRs.

**3. Plasmodium unique genes**

We will focus in one of the most fascinating categories of protein families of *Plasmodium* pan-genome, the unique proteins (i.e. proteins with no significant sequence similarity within the pan-genome). Specifically, we aim in an *in-depth* comparative genomic analysis of the *Plasmodium* pan-genome *unique proteins.* The unique proteins of the *Plasmodium* pan-genome are of utmost interest in our efforts to explore and understand the molecular biology of the malaria parasites. Genes or proteins with no homologs within *Plasmodia* could be:

1. *Orphan/Strain specific* providing vital information regarding *pathogenicity* and be good targets for *drug/vaccine design.*
2. Indicators of *Horizontal Gene Transfer (HGT)* or *Loss/Gain events* from the other *Plasmodium* species and may shed light into *Plasmodium* evolutionary history (i.e. *species specific*).
3. Present in other *Plasmodia* but *have not yet* been annotated.
4. *Gene prediction* or *functional annotation artefacts* and thus, through our careful manual annotation provide the correct sequence/annotation.

5. Could be products of *contamination* where it should be eliminated both from the *Plasmodium* genomes and sequence databases in order to prevent "pollution" of public sequence databases.

We anticipate that this set of experiments will enhance our understanding on the evolutionary forces shaping *Plasmodium* pathogenicity and provide the necessary knowledge and tools for envisage novel and thus, highly effective anti-malarial drugs/vaccines targets.

4. **CBRs structural signatures**

These sets of experiments are directed to determine the *structural preferences* of CBRs based on data retrieved from experimentally solved protein structures. Using a dataset with more than 4000 non-redundant protein sequences from PDB (Berman et al., 2000), we aim to determine clusters of CBRs sharing similar structural preferences by computing the relative solvent accessibility and secondary structure content. The fact that CBR existence is known to create artifacts throughout the analysis pipeline of structure and function determination, further research is needed.

The present thesis aims to investigate the effects of handling CBRs in computing pan-genomes and deploy the acquired knowledge for comparative genomics. Particularly, we propose an optimal strategy for computing pan-genomes and utilize it with malaria parasites. Our results will provide new insights into *Plasmodium* evolutionary history and may pave the way for novel biotechnological/biomedical applications.

Due to the increasing availability of complete sequenced genomes, it is expected that our approach may be extended to a wide range of biological organisms, thus elucidate the biology behind the species under study. For example, it will be feasible to expand our approach to study/predict the physiological and pathological properties of new sequenced strains and their connection with other pan-genomes. Answers to such questions are currently not available and may enlighten our understanding on the evolutionary forces that shape modern genes and genomes.

From a more practical perspective, our results will provide novel knowledge and the capacity to design and build tools enabling cutting-edge applications in the biotechnological and biomedical sector. For instance, it can be envisaged that by determining unique protein families of the *Plasmodium* species may provide clues on

devising more efficient techniques for drug/vaccine development. However, another practical dimension, in the biomedical domain, would be to provide sophisticated tools for predicting the drug resistance of a genome.

Finally, we intend to make all software tools, results and data produced in this study available to the research community, using appropriate – but permissive – licenses, which we hope will (i) act as dissemination of our work, and (ii) enable other research groups to expand this work.

# Chapter 2 – Data and Methods

## 2.1. Introduction

One of the key challenges of modern molecular biology is the increasing number of biological data (e.g. fully sequenced genomes) and the realization that is impractical to analyze it manually. Bioinformatics, as an interdisciplinary field of science, combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Its primary goal is to increase the understanding of biological processes by developing and applying computational techniques. In our days, bioinformatics and computational biology proved to be an important ally of structure/function prediction tools, comparative genomic analyses and pan-genome studies by integrating data from various fields of biology.

In this chapter, we describe the data retrieval process and computational framework followed for achieving our hypothesis and objectives.

## 2.2. Computational Systems and Tools

All computations described in the following were performed on a Linux workstation with Intel (R) Core (TM) 2 Duo CPU E8400@3.00GHz with 4GB RAM, operating under the freely available Ubuntu 16.1 (i586) distribution. All custom programs developed in this study were written in Perl (v5.10.0) or in R scripting language (https://www.r-project.org/; (R Core Team, 2017)) using the open source RStudio software (RStudio Team, 2015).

Perl is a programming language that was developed mostly for easy information extraction and adaptive printing reports. However, Perl is a well-known programming language mostly among Bioinformaticians and Computational Biologists due to developing either small programs (scripts) or fully-fledged software packages.

R is a programming language and environment for statistical and graphical computations, developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues (R Core Team, 2017). RStudio is an integrated development environment (IDE) for R which, it includes tools and packages for statistics, plotting and workspace management (MRAN, 2018; RStudio Team, 2015).

Several external modules were provided as source code bundles, in different programming languages. In particular, for building CAST (Promponas et al., 2000) we compiled the respective source code with gcc version 4.3.2.

## 2.3. Data

### Genome sequences

In this study, we selected only the *Plasmodium* species and strains with a completely sequenced genome (**Table 2**). We downloaded each of the *Plasmodium* genomes in our dataset from the dedicated web-database, PlasmoDB website (C. Aurrecoechea et al., 2009). The comparative genomics analysis pipeline was performed on a subset of the *Plasmodium* species in **Table 2** due to the fact that, at the time of the analysis, there were only nine fully sequenced *Plasmodium* genomes deposited at the PlasmoDB (downloaded on the 02/07/2013; PlasmoDBv9.3; Release Date: 09/07/2012 (C. Aurrecoechea et al., 2009). However, for the pan-genome unique proteins we downloaded all *Plasmodium* species from a newer version of PlasmoDB (downloaded on the 08/02/2017; PlasmoDBv30; Release Date: 26/01/2017) and sequence data were processed following the same methodology as in comparative genomics analysis pipeline.

Sequence data were codified following the style of the COGENT database (Janssen et al., 2003) for consistency and easy manipulation from computational tools. The COGENT style encodes genus and species names into four-character identifier prefix string, followed by a code for the strain name, its version and for proteins the relative order of the sequence within the proteome (Janssen et al., 2003). For example, the *P. falciparum* 3D7 rifin (RIF) protein with a PlasmoDB id PF3D7_0413300 is codified as PFAL-3D7-01-1 (**Figure 5**).

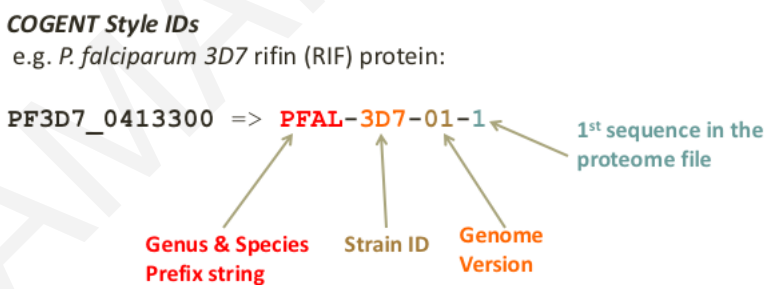



**Figure 5:** An example of how the sequence data were codified using the COGENT style.

Along with the codification of the protein ids we compiled all protein sequence data from the complete sequenced *Plasmodium* species into a single data collection (namely pan-genome file). However, we should note that we excluded all protein sequences with lengths less than 30 residues long, since some of them are questionable (Eberhardt et al., 2012; Höps et al., 2018) . In fact, we eliminated protein sequences with as little as a single amino acid residue. In total, we excluded 599 protein sequences in the nine *Plasmodium*

species dataset and 2,240 protein sequences from the full species dataset. An extra step of sequence elimination was performed in the full species dataset as, in addition to the short sequences, where also eliminated all protein sequences containing the star (*) character. The star (*) character often means the STOP codon indicating low quality coding sequence or suggest that the sequence is a pseudogene (Solovyev et al., 2006). Pseudogenes are not actually proteins but functionless relatives of the genes that have lost either their expression in the cell or the ability to code protein (Solovyev et al., 2006; Vanin, 1985). The final pan-genome files consisted of 50371/108563 protein sequences respectively in FASTA format.

**Table 2:** The nineteen Plasmodium species with complete sequenced genomes and their number of proteins (inside the parenthesis we note the number of proteins after the elimination of the questionable protein sequences), host and pathogenicity (C. Aurrecoechea et al., 2009).

| A/A | Plasmodium Species | Strain | Malaria Type | Host | No. Proteins | Average %AT | Completion Date | Data Source | Reference |
|---|---|---|---|---|---|---|---|---|---|
| **1.** | *P. berghei* | ANKA | Mild | Rodents | 5076 (**4969**) | 74.97 | 03/01/2012 | GeneDB | (Fougère et al., 2017, 2016) |
| **2.** | *P. chabaudi* | Chabaudi | Mild | Rodents | 5217 (**5202**) | 73.34 | 03/01/2011 | GeneDB | (Hall et al., 2005; T. D. Otto et al., 2014; Sato et al., 2013) |
| **3.** | *P. coatneyi* | Hackeri | Mild | Monkeys | 5516 (**5516**) | 58.36 | 22/03/2016 | Genbank | (Chien et al., 2016) |
| **4.** | *P. cynomolgi* | B | Lethal | Monkeys, **Humans** | 5716 (**5714**) | 59.6 | 06/01/2012 | Genbank | (Ta et al., 2014; Tachibana et al., 2012) |
| **5.** | *P. falciparum* | 3D7 | Lethal | **Humans**, African apes | 5548 (**5436**) | 74.94 | 02/01/2012 | GeneDB | (M. J. Gardner et al., 2002; Hall et al., 2002) |
| **6.** | *P. falciparum* | IT | Lethal | **Humans**, African apes | 5480 (**5403**) | 75.13 | 02/01/2012 | GeneDB | Unpublished |
| **7.** | *P. fragile* | Nilgiri | Mild | Monkeys | 5672 (**5671**) | 58.32 | 20/03/2015 | Genbank | (Jongwutiwes et al., 2005) |
| **8.** | *P. inui* | San Antonio1 | Mild | Old World monkeys | 5832 (**5832**) | 57.05 | 31/01/2014 | Genbank | Plasmodium 100 Genomes |
| **9.** | *P. knowlesi* | H | Severe/Non-lethal | **Humans**, Monkeys | 5323 (**5314**) | 60 | 03/01/2012 | GeneDB | (Jongwutiwes et al., 2005; Pain et al., 2008) |
| **10.** | *P. malariae* | UG01 | Severe-Lethal renal complications | **Humans** | 6573 (**6033**) | 71.35 | 03/02/2017 | GeneDB | (Rutledge et al., 2017) |
| **11.** | *P. ovale* | curtisi GH01 | Non-lethal | **Humans** | 7165 (**6705**) | 68.13 | 19/9/2016 | GeneDB | (Rutledge et al., 2017) |
| **12.** | *P. reichenowi* | CDC | Mild | African apes | 5848 (**5729**) | 74.6 | 03/09/2014 | GeneDB | (Thomas D. Otto et al., 2014) |
| **13.** | *P. vinckei* | Petteri CR | Mild | Murine | 5160 (**5160**) | 73.7 | 31/01/2014 | Genbank | Plasmodium 100 Genomes |
| **14.** | *P. vinckei* | Vinckei | Mild | Murine | 4954 | 74.15 | 17/01/2014 | Genbank | Plasmodium 100 Genomes |

| A/A | Plasmodium Species | Strain | Malaria Type | Host | No. Proteins | Average %AT | Completion Date | Data Source | Reference |
|---|---|---|---|---|---|---|---|---|---|
| | | vinckei | | | (**4954**) | | | | |
| **15.** | *P. vivax* | P01 | Severe/Non-Lethal | **Humans** | 6670 (**6554**) | 58.05 | 18/06/2015 | GeneDB | (Auburn et al., 2016) |
| **16.** | *P. vivax* | Sal1 | Severe/Non-Lethal | **Humans** | 5552 (**5534**) | 55.28 | 06/13/2007 | GeneDB | (Carlton et al., 2008; Iwagami et al., 2010; Jongwutiwes et al., 2005) |
| **17.** | *P. yoelii* | 17XNL | Mild | Rodents | 7724 (**7139**) | 74.96 | 09/01/2005 | Genbank | (Hall et al., 2005; T. D. Otto et al., 2014; Vaughan et al., 2008) |
| **18.** | *P. yoelii* | YM | Mild | Rodents | 5685 (**5646**) | 74.86 | 07/01/2012 | GeneDB | (Vaughan et al., 2008) |
| **19.** | *P. yoelii* | 17X | Mild | Rodents | 6092 (**6053**) | 74.92 | 01/02/2014 | GeneDB | (T. D. Otto et al., 2014) |
| **Total:** | | | | | 110803 (**108563**) | | | | |

**Functional annotation**

The functional analysis of proteins with known function of the *Plasmodium* pan-genome and by extension, of their phylum (Apicomplexans) could determine novel potential drug/vaccine targets or evolutionary patterns of this phylum. So far, experimentally determined functional annotation exists only for *P. falciparum* 3D7 or by inference through prediction tools.

For the functional analysis we used information provided by the GO project (T. D. Otto et al., 2014) and protein product description provided by PlasmoDB (C. Aurrecoechea et al., 2009). In details, the GO project is an international collaboration of various databases, including PDB and UniProtKB, which could be described as an attempt for consistent descriptions of gene products among different databases.

The gene ontology and annotation files for the *P. falciparum* were downloaded from the Gene Ontology website((Ashburner et al., 2000), access date: 03/07/2013).

We used the gene ontology file (.obo) which contains the terms, definitions and ontology structure from the GO Ontology downloads website.

We downloaded, also, the filtered gene association file (access date: 03/07/2013) for PDB, for identifying statistically significant representations of functions in groups of sequences sharing the same CBR type.

**Orthologous genes (PlasmoDB)**

A critical step in clustering of the *Plasmodium* CBR-containing proteins is the identification of orthologous genes that either belong or "do not" belong in a given orthologous group. PlasmoDB orthologous groups are generated from the OrthoMCL database (Chen et al., 2006; Li et al., 2003). The OrthoMCL algorithm clusters proteins into orthologous groups based on BLAST similarity across multiple eukaryotic genomes. Thus, the list of orthologous genes was downloaded directly from the PlasmoDB website (http://plasmodb.org/plasmo/) on 03/07/2013.

A custom Perl script was developed for mapping the PlasmoDB protein ids to the COGENT style ids of our dataset and another one for performing the comparisons between our clustering methodology and OrthoMCL orthologous clusters. The purpose of the first script was to identify those OGs that are identical to the ones in our analysis. One issue that we must consider here is the fact that OrthoMCL does

not include *P. cynomolgi* in their datasets. In order to bypass this and be able to find all identical OGs, we ignore *P. cynomolgi* proteins from our analysis. Specifically, we create bit vectors of each cluster and OG and compare them bit-by-bit with the *P. cynomolgi* bit turned off. If both vectors are identical, then we consider them as robust.

Afterwards, we follow the same approach as in MCL clustering analysis (see *Clustering* section).

**Protein Sequence Data**

We selected for our dataset only high-resolution protein structures solved by X-ray crystallography. X-ray crystallography provides a single snapshot (conformation) of the molecule under study enabling the determination of the atoms in the crystal, their chemical bonds, as well as their disorder regions. It is essential to only include protein sequences sharing sequence identity below a certain threshold, to have a unique representative of each protein family, thus removing redundancy.

PISCES is a protein sequence culling server which offers subsets of protein sequences collected from the entire PDB per structure quality and maximum mutual sequence identity (Wang and Dunbrack, 2003). The PISCES website (Wang and Dunbrack, 2003) provide users several options for the creation of non-redundant datasets. Users can use the online culling server for the creation of non-redundant datasets, download pre-compiled non-redundant datasets or download the standalone PISCES package (Wang and Dunbrack, 2003). In this study, we downloaded a pre-compiled dataset in single FASTA-formatted file (access date: 8/11/2016 – last update: 24/10/2016) from PISCES website. Our non-redundant protein sequences dataset, namely PISCES file hereafter, was composed of 4424 protein sequences with less than 30% sequence identity, resolution ≤ 1.6 Å and R-factor ≥ 0.25.

**Protein Structure Data**

For all the protein sequences contained in the protein sequence dataset, we downloaded the respective PDB files, which contained the atomic resolution three dimensional structures (Berman et al., 2000). Along with the non-redundant protein sequence file, we downloaded the respective list of PDB ids. This list was subjected to the batch online service of PDB

(http://www.rcsb.org/pdb/download/download.do; access date: 8/11/2016; (Berman et al., 2000)) for downloading the respective protein structures.

**Calculating the Relative Accessible Surface Area (RASA)**

NACCESS (Hubbard and Thornton, 1993) is a program implementing the method of Lee B. and Richards F.M. (Lee and Richards, 1971), for calculating the atomic accessible surface of a protein structure. This method is based on rolling a probe of given size around the van der Waals surface of atoms in a protein structure (Hubbard and Thornton, 1993). The program takes as input a PDB formatted file and may be fine-tuned by user defined option parameters. In this study, NACCESS was used with the default settings. Upon completion of its computations, NACCESS outputs 3 text files: a log file, one containing the calculated accessible surface for each atom in PDB file and the RASA file containing the summed atomic accessible surface areas over each protein. Additionally, the relative accessibility of each residue is included, calculated as the percentage accessibility compared to the accessibility of that residue in the ALA-x-ALA tripeptides.

**Secondary Structure Elements for the PDB entries**

The Dictionary of protein Secondary Structure Pattern (DSSP; (Kabsch and Sander, 1983)) is a database of secondary structure assignments for all protein entries in PDB. The user can obtain the pre-calculated DSSP files for all PDB entries using the rsync command of the ftp service (Joosten et al., 2011; Kabsch and Sander, 1983). Additionally, the user can download the standalone DSSP program for creating its own DSSP files for a given protein structures dataset.

In this study, we downloaded a pre-calculated DSSP FASTA-formatted file from the PDB website (http://www.rcsb.org/pdb/static.do?p=download/http/index.html#ss; download date: 15/11/2016; (Berman et al., 2000)).

## 2.4. *Computational framework for comparative genomics*

**Data Collection**

All protein sequence data from the nine/nineteen completed *Plasmodium* species were compiled into a single data collection (in FASTA format) using a custom developed tool (Figure 6). In addition, this tool codified all sequence data following

the COGENT database style (Janssen et al., 2003) as described in *Genome Sequences* and provides a list of mappings of the PlasmoDB ids to the COGENT ids.

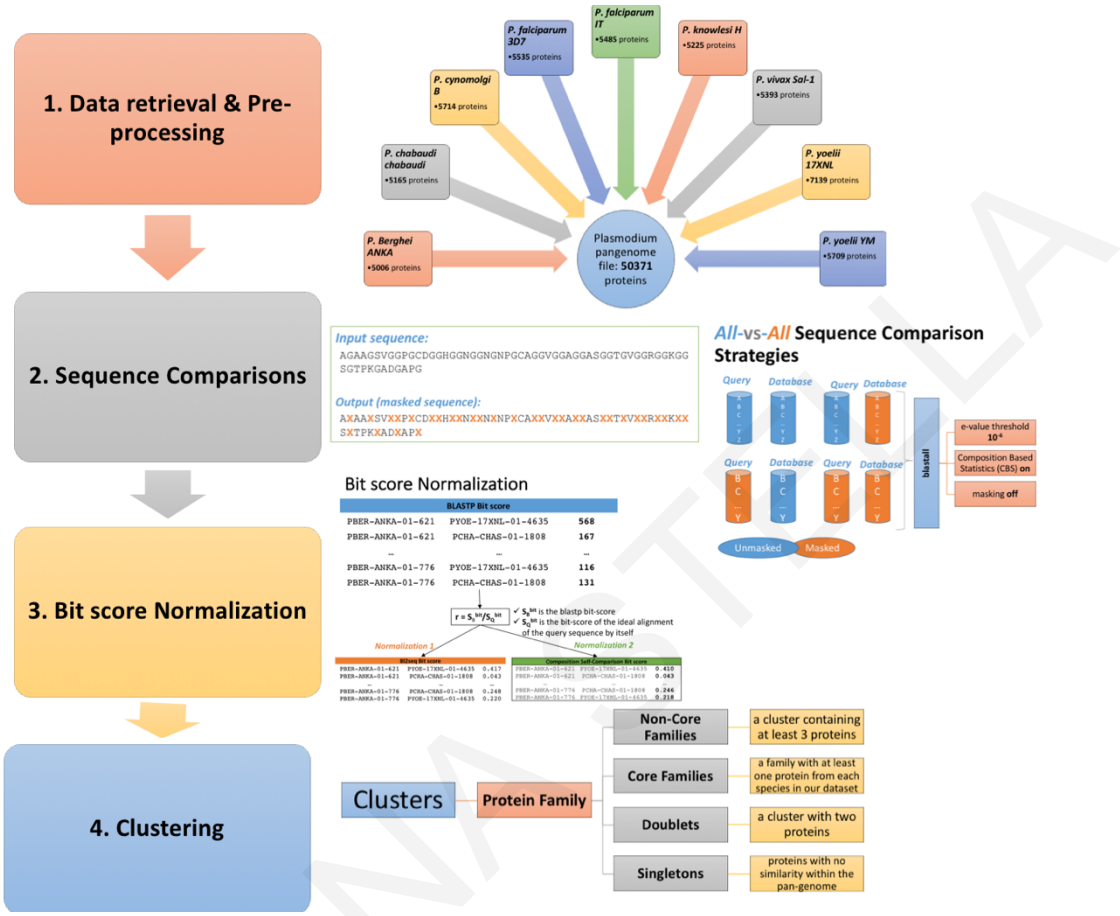



**Figure 6:** A flowchart of the comparative genomics analysis pipeline.

**Detection of CBRs**

Different definitions and, thus, different detection and masking algorithms, have been proposed over the years. In this study, we choose the CBR detection algorithm CAST (Promponas et al., 2000) since earlier studies illustrated the superiority of CAST when applied as a filter prior to BLAST (Altschul et al., 1990) searches and due to its selective detection (and masking) properties (Kirmitzoglou, 2014; Kreil and Ouzounis, 2003; Promponas et al., 2000).

Subsequently the pan-genome file was used as input file for CAST (Promponas et al., 2000) using the default parameters (threshold 40 and BLOSUM62 substitution matrix) for the detection and masking of CBR regions (Figure 6). In total, 32136 CBR proteins and 91327 CBRs were filtered out for further study.

The CBRs of the experimentally solved structures were detected using a different cut-off value for CAST. Specifically, the PISCES dataset was used as the required input file of CAST (Promponas et al., 2000) where, we selected as cut-off value 25 and the BLOSUM62 substitution matrix. CAST mode 25 has been shown to be the most appropriate CBR detection mode when seeking structural features of CBRs (Tamana et al., 2012).

**Sequence Comparison**

Afterwards, both the unmasked and masked pan-genome files were transformed to BLAST-able database files using the standalone BLAST suite of tools (Altschul et al., 1990) (Figure 6). We investigate the effect of CBR-masking on the elucidation of the Plasmodia pan-genome structure following an *all-against-all* approach (blastall: e-value threshold $10^{-6}$, Compositional Based Statistics on, masking mode turned off). Specifically, we performed 4 different sets of experiments by masking both the query and database files or masking either query/database file or no masking either file (Figure 7).

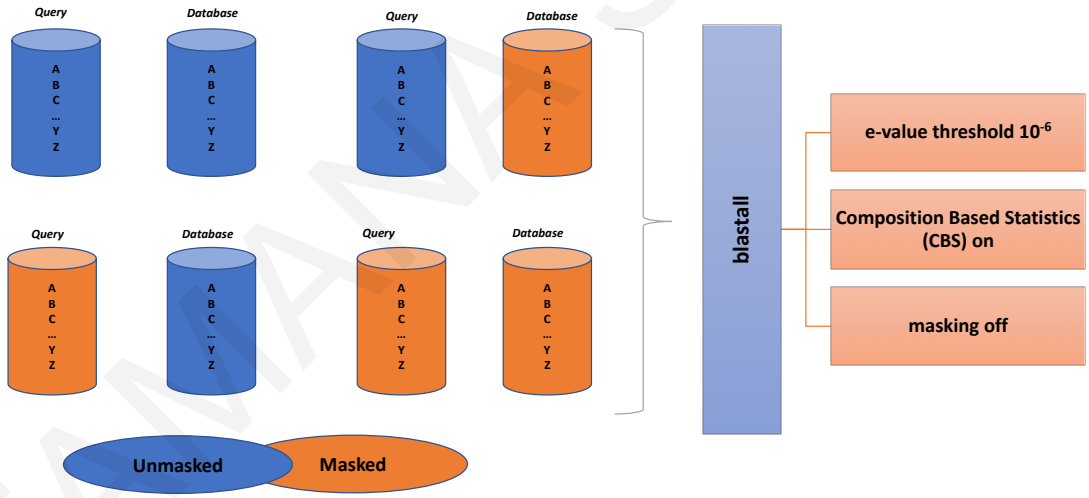



**Figure 7:** A schematic representation of the exhaustive all-vs-all the sequence comparison strategies we performed.

In order to test for possible artifacts, along with the BLAST results, we calculated the normalized bit score (namely R-fraction, see equation 1):

$$R = S_B^{bit} / S_Q^{bit} \qquad (1)$$

where $S_B^{bit}$ is the reported bit-score and $S_Q^{bit}$ is the bit-score of the ideal alignment of the query sequence by itself. Bit-scores of the self-alignment can be obtained either by employing the blast2seq (Altschul et al., 1990) program for performing the self-comparison or by explicitly using the query composition and the BLOSUM62 scoring matrix. For this purpose, a batch script was developed that (i) fed each protein sequence in the pan-genome file as the required by the blast2seq tool query and database sequence and execute the blast2seq using e-value cutoff $1e^{-3}$, (ii) called a function to calculate the sequence's composition and (iii) calculate the R-fraction of each method's output respectively.

**Clustering**

The exhaustive *all-against-all* comparisons are then subject to MCL clustering (Enright et al., 2002; Van Dongen, 2000) for identifying protein families in Plasmodia (Table 3). The MCL algorithm finds clusters by representing sequence similarities as a connection graph. It relies on Markov matrices and in two operators for the assignment of proteins into families based on pre-computed sequence similarity information (Enright et al., 2002; Van Dongen, 2000).

MCL was employed with default parameters and results were stored for further analysis. This procedure is performed based on the standard bit-scores reported by BLAST.

**Table 3:** List of all the MCL runs we performed where **green colored cells** denote usage of the respective mode whereas empty cells denote absent mode. The Bl2seq tool was employed using cutoff e-value: $1e^{-3}$ while the Composition Self Comparisons computed using BLOSUM62 substitution matrix. **Column abbreviations: DB:** data base file, **QU:** query file and **CSC:** Composition Self Comparisons.

| MCL ID | All-vs-All Masking | | Bit Score Normalization | | | | MCL ID | All-vs-All Masking | | Bit Score Normalization | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DB | QU | Mode | | Masking | | | DB | QU | Mode | | Masking | |
| | | | Bl2seq | CSC | DB | QU | | | | Bl2seq | CSC | DB | Query |
| MCL1 | | | | | | | MCL11 | | | | | | |
| MCL2 | | | | | | | MCL12 | | | | | | |
| MCL3 | | | | | | | MCL13 | | | | | | |
| MCL4 | | | | | | | MCL14 | | | | | | |
| MCL5 | | | | | | | MCL15 | | | | | | |
| MCL6 | | | | | | | MCL16 | | | | | | |
| MCL7 | | | | | | | MCL17 | | | | | | |
| MCL8 | | | | | | | MCL18 | | | | | | |
| MCL9 | | | | | | | MCL19 | | | | | | |
| MCL10 | | | | | | | MCL20 | | | | | | |

**Statistical analysis**

A specialized module was developed for calculating descriptive statistics for testing the effects of CBRs in computing the Plasmodia pan-genome. This module calculated the average and median sequence length of the clusters for each MCL run and the Wilcoxon Rank Sum test.

## 2.5. Unique genes in malaria parasites: a pan-genomic approach

**Detection of Plasmodium pan-genome unique proteins**

From the clustering file, we extracted 1201 unique proteins and performed a sequence similarity search against a masked version of the NR database (e-value threshold: $1e^{-6}$, Composition Based Statistics (CBS) off, filtering off). In details, we downloaded the NR database from the NCBI ftp site (ftp://ftp.ncbi.nlm.gov/blast/db/FASTA/; download date: 14/2/2017) which, at the time, composed of 114,103,265 non-redundant protein sequences. Subsequently, the NR database was set as input to CAST (Promponas et al., 2000) using the default parameters (threshold 40 and BLOSUM62 substitution matrix) for the detection and masking of CBRs. The masked NR database was transformed to BLAST-able database file using the standalone BLAST suite of tools (Altschul et al., 1990) and fed to BLASTP for the *all-vs-all* sequence comparisons.

**Analysis of the Plasmodium unique proteins against NR/NT database**

The resulting sequence comparison file, using a custom-made tool, was divided into two categories: the putative de novo proteins (i.e. proteins that did not find any match against the NR database or their only match was itself) and the multiple hits proteins (>= 2 hits). Subsequently, the multiple hits files were subjected into efetch utility program provided by NCBI (Sayers and Miller, 2010) as to fetch the FASTA sequence of each hit. The efetch program requires a list of UIDs or Accession numbers separated by comma and returns their sequence in FASTA format (for more details see the NCBI's e-utilities instruction manual (Sayers and Miller, 2010). The putative de novo FASTA protein sequences were retrieved directly from the unique protein's file.

All putative unique proteins were, then, subjected to the standalone version of

tBLASTN (e-value: $1e^{-6}$, CBS off, filtering off) for the identification of nucleotide sequences encoding proteins like the query. Specifically, we downloaded the NT database from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/; download date: 20/2/2017) which, at the time, composed of 41,308,655 sequences. The resulting output file was stored for further process by a custom-made tool.

Finally, for the multiple hits (>= 6 hits) we enlisted the standalone version of mview (Li et al., 2015) for constructing Multiple Sequence Alignments (MSAs) from the BLASTP pairwise alignments. Specifically, for each unique protein in this category we formatted its hits file as a BLASTP database and re-run BLASTP using the respective protein as the query sequence. The pairwise alignments were set as the required input file to mview, for the reconstruction of MSAs. This approach enables local alignments of the sequences instead of the global alignments that calculated from ClustalW (Sievers et al., 2014). Local alignments provide useful information in cases of dissimilar sequences but suspected to share a domain or motif.

**Scoring measures for detecting genuine unique proteins**

For each genome in our dataset we calculated four selective criteria (length, %CBR, %G+C and Neighborhood Distribution) based on the proposed methodology of (Wilson et al., 2007). Wilson and colleagues developed a scoring method, called 'Quality Index for Predicted Proteins (QIPP)', for ranking orphans and TRGs in bacterial and archaeal genomes. Their selected criteria are length (Skovgaard et al., 2001), percentage low complexity (%CBR, (Altschul et al., 1994), difference in G+C composition of sequence and genome (%GC (Navarre et al., 2006), average amino acid cost (Akashi and Gojobori, 2002; Heizer et al., 2006) and neighborhood distribution (ND, (Zheng et al., 2005)).

In this study, we selected to calculate only the scores for length, %CBR, %GC and ND criteria since calculating the amino acid biosynthesis cost for *Plasmodium* is limited to plasma free amino acids and host cell heamoglobin and thus, is induced with the extra effort of determining which are the essential amino acids (that each species could not find in their respective hosts) for each *Plasmodium* species in our dataset.

For each genome and for each of the four selected criteria, the distribution of non-orphans was generated and the percentiles for that distribution were calculated

using R studio (RStudio Team, 2015). Length was calculated as the total number of amino acids and was transformed into a sub-score from 0-100 depending on the percentile in which it fell. The percentage CBRs (%CBR) was calculated as the number of masked residues versus the total number of amino acids in sequence and the score was subtracted from 100. Percentage G+C (%GC) for each sequence was calculated from its respective CDS (downloaded from the PlasmoDB website using the 'Sequence Retrieval' tool) and the score was calculated as the deviation from the mean value. Values above the 50$^{th}$ percentile were corrected by the equation: 100 minus the percentile value multiplied by two while, values below the 50$^{th}$ percentile had their percentile doubled. ND was calculated by determining the level of conservation of the five flanking CDS on either side of a unique CDS. In details, for each genome in our dataset, first, we separated the proteome into their respective contig/chromosome based on the genomic location provided in the header of each protein. Then, using the pan-genome BLASTP output file we computed the pan-genome Best Bidirectional Hits (BBHs). Finally, for each unique protein we counted the number of proteins with BBHs in at least half plus one (i.e. 11) other *Plasmodium* species in a neighborhood of -/+5 proteins (i.e. maximum of 10 proteins should be recorded). The final number was divided by the number of conserved proteins in the flanking region and was transformed into a sub-score from 0-100 depending on the percentile in which it fell. For obtaining a final QIPP score between 0 and 1, the average is taken and divided by 100. A zero QIPP score will indicate the worst possible candidate for a real gene and one would be strong indication for a real gene. Because malaria parasites are known for their heavily biased genomes the %CBRs criterion might skew the final score, we also calculated QIPP scores without the %CBRs score.

**Unique versus pan-genome amino acid/codon usage statistics**

Custom made PERL scripts were developed for calculating the amino acid (masked and unmasked) and AT-rich codon usage distributions, both of the putative unique (1201 protein/CDS sequences) and pan-genome proteins (108563 protein/CDS sequences). Specifically, for the amino acid distributions, using the putative unique protein ids and the pan-genome protein sequence file, we first calculate the amino

acid distributions (i.e. the average number of each residue) in both datasets. It's worth mentioning that for the pan-genome distributions (both for amino acids and codon usage) we excluded from calculations the unique protein sequences and thus, avoiding calculating duplicated values.

Then, for each residue type we computed the log ration of the pan-genome versus unique proteins based on the following formula:

$$\boldsymbol{x_{AA}} = \log \boldsymbol{x_{AA_{pan-genome}}} / \boldsymbol{x_{AA_{unique}}} \ (\mathbf{2})$$

The AT-rich codon usage distribution was calculated following a similar approach to the amino acid distribution (i.e. average number of each codon in our datasets) using each *Plasmodium* species CDS file (retrieved as described in QIPP scores calculations). For the calculations of the AT-rich codons we divided the AT-rich codons in three categories: 1. All Adenine or Thymine (i.e. AAA or TTT), 2. Adenine or Thymine in first and second position (i.e. AAX or TTX or ATX or TAX where X cannot be A or T) and 3. Adenine or Thymine only in first position (i.e. AXX or TXX where X cannot be A or T).

Ambiguous codons (e.g. AMW or YTT) were eliminated and not considered for the final calculations. In total, 21,535 ambiguous codons out of 72,739,511 codons were excluded in the pan-genome dataset while, in the putative unique dataset we eliminated 45 ambiguous codons out of a total of 125,003 codons.

Finally, for both datasets and for each category, we calculated the same log ration (see equation **2**) as for the amino acid distribution.

**Isolation Index of Organisms**

The degree of Isolation of a given genome can be quantified using the *Isolation Index of Organisms* (IIO)(Fukuchi and Nishikawa, 2004). The IIO is based on the e-value obtained by BLAST searches. Here, we computed IIO for every species in our dataset by computing the average logarithm of the e-values of BBHs. When an e-value was zero, we set a relatively small e-value ($1e^{-175}$; if the reciprocal best hit it was not itself) or the e-value was set as the cut-off threshold used in BLAST search ($1e^{-06}$; for the unique genes). Paralogous genes that were not unique (had also hits in other *Plasmodium* species) were discarded. We choose to calculate IIO based on reciprocal best hits e-values instead of just using the best hit e-value as in (Fukuchi and

Nishikawa, 2004) because reciprocal best hits by definition provide further evidence of the sequence similarity.

**Plasmodium core genome phylogenetic tree**

The *Plasmodium* core genome phylogenetic tree was constructed based on core cluster protein families (i.e. single copy proteins found in all *Plasmodium* species in our dataset) using mrBayes program (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) on CIPRES Science Gateway(Miller et al., 2010).

First, we downloaded all protein sequences of the avian-host malaria parasite *P. gallinaceum* (download date: 7/4/2017; https://www.ncbi.nlm.nih.gov/protein/?term=Plasmodium+gallinaceum) and *P. relictum* (download date: 7/4/2017; https://www.ncbi.nlm.nih.gov/protein/?term= Plasmodium+relictum) in FASTA format from NCBI website. Then, we re-constructed the *Plasmodium* pan-genome and protein families by following the same analysis as before. In total, we retrieved 1895 *Plasmodium* core clusters. As outgroup species we selected *Toxoplasma gondii* ME49 (download date: 30/1/2017; from ToxoDB_v29 (Gajria et al., 2008); release date: 12/10/2016) because it belongs in the phylum of Apicomplexa and closely related to *Plasmodium* species (Reid et al., 2012). The *T. gondii* protein sequences were post-processed into COGENT ids for easy manipulation from computational tools and compared against *Plasmodium* core clusters using the standalone version of HMMER (Finn et al., 2015) . This step enables to retrieve only the *Plasmodium* core clusters that are composed of protein sequences that share statistically significant sequence similarity to *T. gondii* proteins leading to a final dataset of 1700 *Plasmodium* plus *Toxoplasma* core clusters.

Consequently, for each cluster we performed MSAs using the standalone version of CLUSTAL Omega (Sievers et al., 2014). The individual MSAs were, then, merged into one super matrix MSA in FASTA format composed be 1342472 aligned positions. The super matrix MSA was converted into the required by mrBayes NEXUS format and set as input file for mrBayes on CIPRESS Science Gateway (Miller et al., 2010).

## 2.6. OST complex subunits in protists

### Keyword-based literature and database search

Initial search of the biomedical literature was performed between March-May 2017 as described in the following and returned no results reporting the presence of any of the genes/proteins in question in publications co-mentioning the term 'plasmodium'. Example searches of this kind in the PubMed® database involved queries such as '(ost4 or ost4p) AND plasmodium'. Gene synonyms were retrieved for yeast and human OST subunits in order to make the search comprehensive. Since PubMed® searches are performed only against publication abstracts (as opposed to full-text publications), we also performed searches against the free full-text archive in PubMed Central® (https://www.ncbi.nlm.nih.gov/pmc/). In addition, we manually verified entries from citation databases (https://scholar.google.com/; https://www.scopus.com/, http://apps.webofknowledge.com) for citations to the review article by (Kelleher and Gilmore, 2006) that may possibly report the presence of any of the aforementioned subunits in *Plasmodium*. All relevant publications were manually verified that no mention of the aforementioned OST subunits in *Plasmodium* has been reported.

PlasmoDB (http://plasmodb.org), collects a wealth of genome and functional data for *Plasmodium* species from multiple resources (C. Aurrecoechea et al., 2009). While this work was performed (October 2017), 22 essentially complete *Plasmodium* genomes were available through PlasmoDB. This resource offers an array of search tools, and it is possible to search via keyword, gene identifier or sequence (e.g. using the NCBI BLAST suite of programs) to retrieve entries of interest. Similar resources to PlasmoDB exist for other protists under the Eukaryotic Pathogen Genomics Resource EuPathDB (Aurrecoechea et al., 2017), of which PlasmoDB is a component database, offering a unified web interface for retrieving information. In particular AmoebaDB (Aurrecoechea et al., 2011), CryptoDb (Heiges et al., 2006) and TrichDb (Cristina Aurrecoechea et al., 2009) were used to retrieve OST subunit information from Entamoeba histolytica, Cryptosporidium parvum, and Trichomonas vaginalis respectively.

In all performed BLAST queries reported herein, we have used the default options provided by PlasmoDB (except for setting a more permissive cutoff for the e-value

equal to 100) and the results were manually inspected to keep only hits with biologically meaningful similarities. In particular, for tBLASTN searches, we either selected all significant hits (matches with e-value <0.001) or –in the case of insignificant hits– we chose to further examine only those genomic regions providing maximum coverage with respect to the query polypeptide prioritizing hits in the syntenic regions examined. The relaxed approach to e-value thresholds is also supported by the small size of the underlying databases.

**Gene prediction**

Gene predictions were performed using the online version of FGENESH+ tool ((Solovyev et al., 2006); version 2.6; http://www.softberry.com/; last accessed June 27, 2018) using the *Plasmodium falciparum* model parameters. Specifically, for each initial tBLASTN hit we retrieved the respective coordinates from the PlasmoDB integrated genome browser tool. Then, depending on the length of the intergenic region between the closest up- and down-stream genes annotated in PlasmoDB, we downloaded 1kbp- or 2kbp-long genomic sequences containing the sequence similarity hit. These sequences were provided as input to FGENESH+ along with an appropriate *Plasmodium* protein sequence as reference, to perform protein-based similarity gene predictions. For *Plasmodium* species where no reliable gene predictions could be obtained using this reference sequence, we set as the required protein-template a protein sequence from a more closely related *Plasmodium* species.

**Syntenic Neighborhood Conservation Index**

A Syntenic Neighborhood Conservation Index (SNCI) is calculated by determining the level of conservation of five flanking genes upstream and downstream of each detected gene possibly encoding a novel OST subunit. In particular, for each of the newly detected OST subunits we retrieve the synteny map from the PlasmoDB website for the respective locus and manually counted the number of conserved genes in all other *Plasmodium* genomes. The number of conserved proteins is then normalized in the 0-1 range by dividing it with the number of the totally considered proteins (maximum 10) to yield the SNCI.

**Multiple Sequence Alignment, Phylogenetic and Structural Analysis**

Multiple sequence alignments for Ost4 subunits were produced using Tcoffee (Notredame et al., 2000) using default parameters as provided by JABAWS (Troshin et al., 2011) via the JalView multiple sequence alignment editor and analysis workbench (Waterhouse et al., 2009). In the case of the longer Ost3/6 subunits, multiple sequence alignments were produced by aligning individual sequences to the OST3_OST6 profile Hidden Markov Model derived from PFAM (PF04756.13) using the hmmalign tool of the HMMER3 package (version 3.2.1, obtained from hmmer.org; (Eddy, 2011)).

A neighbor joining phylogenetic tree for Ost4 subunits was produced using Archaeopterix (version 0.9901 beta; (Han and Zmasek, 2009)), using Poisson correction and was rooted by the mid-point root method.

Geometry analysis of the Ost4 transmebrane helix was performed using HELANAL-Plus (Kumar and Bansal, 2012) using local helix origin points with α-helix definition based on DSSP (Kabsch and Sander, 1983).

## 2.7. Sequence and Structural signatures of CBRs

**Mapping the Relative Accessibility and DSSP patterns to Protein Sequences**

An in-house developed tool was used to map each protein sequence residue to a residue in the respective protein 3D protein structures. In general, protein sequences from PDB entries do not have a 1-1 correspondence to those deposited in protein sequence databases (Djinovic-Carugo and Carugo, 2015). This fact originates from (i) intentional genetic manipulations, used for enabling structure determination (e.g. addition of his-tags, constructs lacking putative aggregation-prone domains) or to understand the principles of protein folding and stability and (ii) disordered regions which are not included in the structure (Djinovic-Carugo and Carugo, 2015). It was originally developed to process each protein sequence from our respective dataset for identifying CBR's burial/exposure patterns. Even though, this study does not address explicitly CBR's burial/exposure patterns, the clever and complicate procedure performed in this script enables correctly correlating the CBR and RASA/DSSP properties for all residues in the dataset. One of the important features

of this script is that it identifies and corrects his-tag containing protein sequences that were wrongly filtered as H-rich by CAST.

The other important feature is the processing of all structures (i.e. the ATOM section of the PDB file) to match the PISCES sequences, thus eliminate as much as possible the problematic PDB entries. From our experience, there are some problematic PDB entries with missing residues or inconsistent numbering. The following example synopsizes major cases describing these types of PDB entries (**Figure 8**).

```
ATOM    2    CA   VAL   A    3     17.997  38.577  42.705  1.00  41.40   C
ATOM    9    CA   ILE   A    4     19.568  41.805  44.109  1.00  24.08   C
ATOM   17    CA   ASN   A    5     20.308  43.996  41.102  1.00  19.67   C
ATOM   43    CA   ASP   A    8     20.096  50.716  46.322  1.00  15.99   C
  …
ATOM  231    CA   GLN   A   31     32.806  34.074  28.628  1.00  12.83   C
ATOM  245    CA   ALA   A   31A    33.842  30.617  29.829  1.00  14.36   C
  …
ATOM  861    CA   ILE   A  109     25.208  57.445  45.913  1.00  10.96   C
ATOM  869    CA   ARG   A  110     27.033  56.297  49.061  1.00  12.73   C
```

**Figure 8:** Case of residue numbering issues. We illustrate cases of *missing terminal residues* from the N- and/or C-terminus, *non–continuous numbering* and, *insertion code* (iCode).

Specifically, this artificial PDB structure misses' residues from the N-terminal that may exist in its primary sequence. Another case of numbering inconsistency appears for the residues ASN5 and ASP8, where the numbering is not sequential. This case may occur either due to disordered atoms not appearing in the structure or when depositors intentionally number continuous residues not as being such, for the sequence positions to be numbered identically to reference sequences. Notably, an insertion code (e.g. ALA31A) is sometimes used in sequence numbering when a protein structure has fewer/extra residues at various places within the chain compared to reference sequences (http://www.wwpdb.org/docs.html). Importantly, there is no way to figure if the most C-terminal residue of the structure is the last residue on the sequence, and there are also cases where N-terminal residues have negative position numbering.

In order to overcome the above-mentioned issues, a Regular Expression (RE) pattern based on the sequence derived from the structural data and possible numbering inconsistencies is constructed. A RE pattern allows fast and flexible string handling

(Schwartz et al., 2005), offering a compressed and generic representation of a sequence or sequence family. REs have been widely used in computational molecular biology to represent sequence families. A very well-known application of REs is the patterns provided by PROSITE (http://expasy.org/prosite/).

In our case, we build a RE of the form:

^(.*)----SEQblock----.{n1}----SEQblock-----(.*)$

to match a structure derived sequence with a 'jump' of *n1* residues in its numbering. We define an *exact match* the case where the PISCES sequence is identical to the structure-derived sequence along their complete lengths. In case of an exact match, then the script prints a table with all four sequences (NACCESS, DSSP, PISCES and CAST) in a tab-delimited text file (Figure 9).

> 1ABCA

| N_AA | DSSP_P | P_AA | LCR_AA | AA_Pos | Fake_AA_Pos | RSA_Value | RSA_Mask | P_Mask | LCR_Mask |
|---|---|---|---|---|---|---|---|---|---|
| R | T | R | R | 1 | 1 | 84.8 | 1 | 0 | 0 |
| E | T | E | E | 2 | 2 | 43.6 | 1 | 0 | 0 |
| A | S | A | A | 3 | 3 | 106.7 | 1 | 0 | 0 |
| … | | | | | | | | | |
| K | S | K | X | 11 | 10 | 55.5 | 1 | 0 | 1 |
| K | S | K | X | 12 | 11 | 57.8 | 1 | 0 | 1 |
| … | | | | | | | | | |
| K | - | K | K | 203 | 175 | 55.6 | 1 | 0 | 0 |
| - | - | K | K | 204 | 192 | 89.1 | 1 | 0 | 0 |

**Figure 9:** An example of the tab-delimited output text file depicting cases of an **exact RE match**, **masked** and **missing/disorder** residues.

When an exact match is not found, then the PISCES sequence is further processed to identify whether it matches with the respective structure-derived RE. With the structure-derived RE we can effectively identify cases of missing residues from the N- or C-terminal, as well as internal residues of the structure, absent due to structural disorder. Of course, our approach depends on the consistency of numbering with missing residues followed by PDB entries. Since this is not always the case, those entries that do not match the RE are finally discarded. When missing residues are identified correctly in the structure derived sequence, they are substituted with a 'placeholder' character, i.e. "-" (Figure 9). For display purposes, we chose to follow an arbitrary position numbering (column 'Fake_AA_Pos'), even when an exact match is observed.

Thus, for these reasons we excluded 83 protein structures from further analysis leading to a total of 4331 protein structures. The final output of this script is a tab-

delimited text file containing a summary table for each protein in our dataset (Figure 9).

**Mapping Structural features to CBRs**

The next step is mapping all structural features of CBRs into a single table for determining clusters based on their structural preferences. Here, we developed a special script that calculates the average RASA and DSSP values for each CBR detected by CAST (Promponas et al., 2000). The DSSP values were divided into the eight classes provided in DSSP website (helix (H), β-bridge (B), extended (E), $3_{10}$-helix (G), π -helix (I), turn (T), bend (S) and loop-irregular structure (L)) plus disorder (D) for missing structural residues.

The script takes as input files the CAST stat file, the RASA/DSSP table, a list of all the his-tag containing proteins (for excluding them from calculations) and the PISCES file. We should note that we did not excluded all his-tags containing proteins but only those that the H-rich CBR was in the first/last 20 residues and the protein did not have other CBRs, leaving only true H-rich CBRs. The output is a table composed of all CBRs and their average RASA/DSSPs.

**CBRs Structural features clustering**

The resulting table was inserted as the required input for the k-means clustering algorithm. The clusters were calculated using R's k-means (RStudio Team, 2015) module with k = 4. The appropriate number of centers was selected after we calculated a plot (**Figure 10**) of the within groups sum of squares by number of clusters extracted also known as Elbow plot (RStudio Team, 2015).

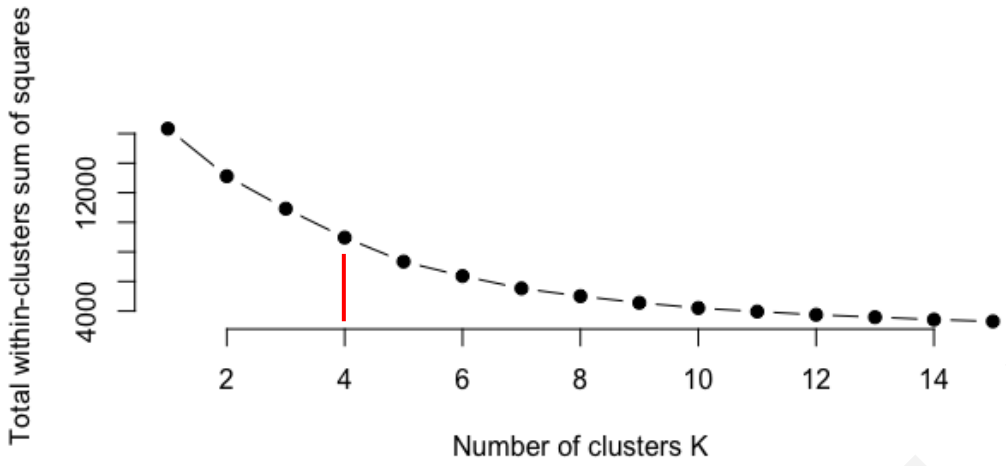



**Figure 10:** A plot of the within groups sums of squares by number of clusters extracted. The **vertical** line indicates our selected number of clusters k.

**Fisher's Test**

Fisher's test is a statistical significance test and is useful for categorical data that result from classifying objects in two different ways (Fisher, 1992). Specifically, it can be used to examine the significance of the association (contingency) between the two kinds of classification. In our case, Fisher's test will help us to identify statistically significant under/over-represented CBRs in each structural feature derived cluster. Ultimately, we will be able to distinguish CBRs favorable structural features and design x-rich (where x is any of the twenty standard amino acids) structural signatures. Thus, our null hypothesis is that the relative proportions of in-cluster x-rich CBRs are independent of the relative proportions of the out-of-cluster x-rich CBRs.

First, for each x-rich CBR in each of the structural features cluster (see Mapping Structural features to CBRs section) we computed 2x2 contingency tables (**Table 4**) and then we computed the hypergeometric distribution as defined by Fisher's test (Fisher, 1992, 1922). The hypergeometric distribution calculates the exact probability of observing the arrangement of the x-rich CBRs inside a specific structural features cluster. For all results reported, we have set a relatively strict significance level of 0.001.

**Table 4:** An example of a contingency table used for the calculations of the hypergeometric test. Abbreviated columns: **w-x-rich:** with the x-rich CBR and **wout-x-rich:** without the x-rich CBR.

| | **w-x-rich** | **wout-x-rich** | **Rows Total** |
|---|---|---|---|
| **In-cluster** | $x_1$ | $y_1$ | $x_1 + y_1$ |
| **Out-cluster** | $x_2$ | $y_2$ | $x_2 + y_2$ |
| **Columns Total** | $x_1 + x_2$ | $y_1 + y_2$ | $\mathbf{x_1 + x_2 + y_1 + y_2}$ |

**Calculating Sequence features of CBRs**

For the sequence features of CBRs we computed the Local complexity $K_1$ and Shannon Entropy $K_2$ (Wootton and Federhen, 1993). The definition of local complexity used here is a combinatorial measure based on the work of Konopka & Owens (Konopka and Owens, 1990; Wootton and Federhen, 1993). Briefly, assuming an amino acid sequence segment of length L, with $n_i$ being the number of occurrences of residue type i complexity $K_1$ is defined as:

$$\boldsymbol{K_1 = \frac{1}{L} log\left(\frac{L!}{\prod_{i=1}^{N} n!}\right)} \qquad \textbf{(3)}$$

where $n! = \prod_{i=1}^{N} i$, with 0! =1 and $\sum_{i=1}^{N} n_i = L$. Thus, complexity $K_1$ can be described as a measure that separates "low" from "highly" biased segments signaling well defined properties of CBRs. Zero or close to zero $K_1$ values should suggest homopolymeric or single-residue repeating regions while $K_1 > 1$ more complex CBRs. In particular, in the trivial case of a homopolymer (i.e. a segment composed of a single amino acid type) $K_1$=0.

Shannon entropy $K_2$ is used as an additional information measure defined as:

$$\boldsymbol{K_2 = -\sum_{i=1}^{N} \frac{n_i}{L} log\left(\frac{n_i}{L}\right)} \qquad \textbf{(4)}$$

and is consider as the expected information content of a protein sequence segment (Wootton and Federhen, 1993).

In addition to SEG-like complexity measures for each CBR we, also, choose to calculate relevant physicochemical properties of sequence segments such as,

hydrophobicity, charge and net charge (all normalized for sequence length), quantities that have been proven useful in determining intrinsically disordered proteins/regions (Uversky et al., 2000). For this purpose, we have opted to use the experimentally derived hydrophobicity scale presented in (Hessa et al., 2005) (available in https://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/defineattrib/hhHydrophobicity.txt; Table 5), while residue charges are simply +1 (R, K), -1 (D, E) and +0.1 (H) (see (Bhagavan and Ha, 2011) for the choice of the partial charge of histidine).

**Table 5**: Experimentally derived hydrophobicity values of the twenty standard amino acids (Hessa et al., 2005).

| Amino Acid | Hydrophobicity value | Amino Acid | Hydrophobicity value |
|---|---|---|---|
| D | 3.49 | **T** | 0.52 |
| E | 2.68 | **C** | -0.13 |
| N | 2.05 | **M** | -0.10 |
| Q | 2.36 | **A** | 0.11 |
| K | 2.71 | **V** | -0.31 |
| R | 2.58 | **I** | -0.60 |
| H | 2.06 | **L** | -0.55 |
| G | 0.74 | **F** | -0.32 |
| P | 2.23 | **W** | 0.30 |
| S | 0.84 | **Y** | 0.68 |

All the above-mentioned calculations have been performed for all the CBRs detected by CAST using a custom-made Perl script (script in the Appendix IV).

**CBRs sequence features clustering**

Following the structural-features clustering we performed k-means clustering using only the sequence features of CBRs in each structural-features derived cluster. We followed a similar methodology, as in structural features clustering, by first determining the optimal k value for each cluster using elbow plots (Figure 11) followed by computing the k-means algorithm in each cluster's sequence feature. We should note that for these sets of clustering we selected to use only Shannon entropy, (normalized) hydrophobicity, charge and net charge since it is known that there exists a strong correlation between Shannon entropy and local complexity and Shannon entropy is much more efficiently computed.

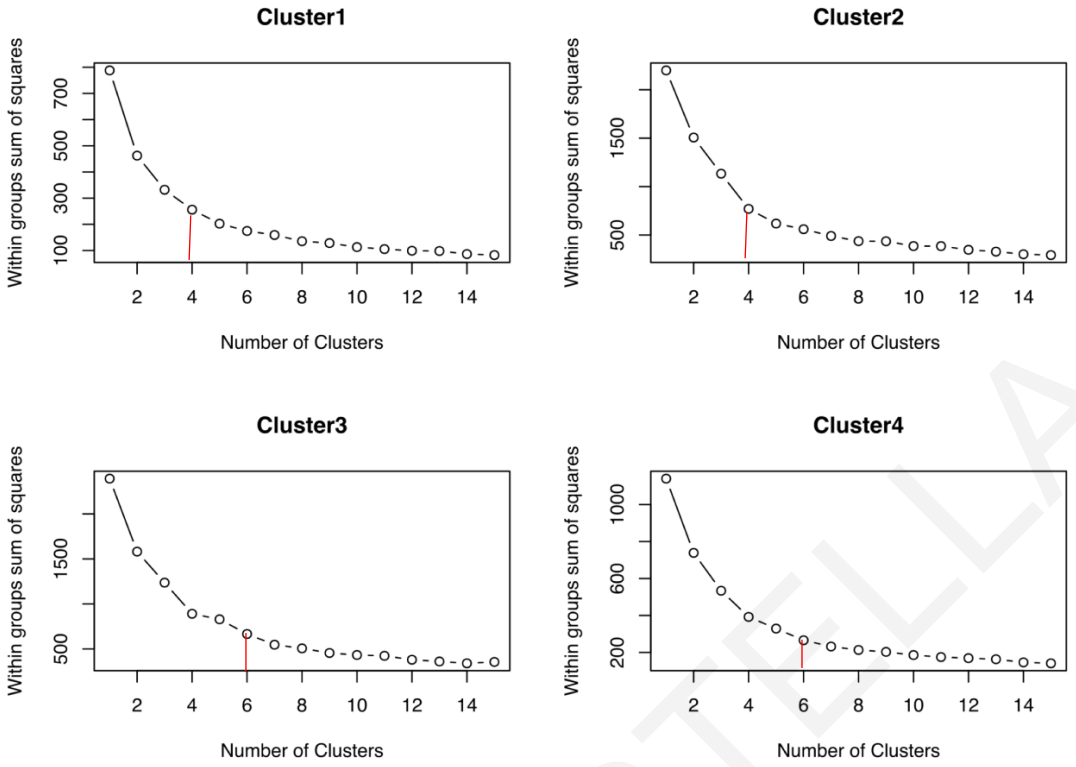



**Figure 11:** A set of the within groups sum of squares by number of clusters extracted (also known as Elbow plot) plots for each structural cluster for determining the optimal number k for the k-means algorithms. The vertical line indicates our selected number of clusters k where for Cluster1 and 2 we choose k=4 and for Cluster3 and 4 k=6.

### Non-parametric and post-hoc statistical tests

We developed custom R code using the RStudio software (RStudio Team, 2015) for computing descriptive statistics, performing visualization and executing appropriate statistical tests to evaluate any statistical differences emerging between the discrete clusters formed.

The Kruskal–Wallis test is a non-parametric method for testing whether three or more independent samples originate from the same distribution (Spurrier, 2003). A significant Kruskal–Wallis test indicates that at least one sample predominates one other sample, and Dunn's test is often used as a post-hoc procedure following rejection of a Kruskal–Wallis test (Dunn, 1964, 1961). Dunn's test is, also, a non-parametric pairwise multiple comparisons formula based on rank sums (Dunn, 1964, 1961). In this study, we formulate our null hypothesis as that the medians of mean hydrophobicity/net charge values of all clusters are equal, and the alternative hypothesis as that at least one population median of one group is different from the population median of at least one other group.

A boxplot is a method for graphically depicting groups of numerical data through their quartiles where the whiskers (lines extending vertically from the boxes) indicating variability outside the upper and lower quartiles and outliers are plotted as individual points (Krzywinski and Altman, 2014). Since boxplots are non-parametric, they display variation in a statistical population without making any assumptions of the underlying statistical distribution and are useful for comparing distributions between several groups or sets of data (McGill et al., 1978). Here, we constructed boxplots to unveil differences between the sequence derived features of CBRs.

# Chapter 3 – Results

## 3.1. A novel framework for pan-genome analysis

In this study, we aim to investigate how CBRs affect the computations of pan-genomes for comparative genomics, utilizing it with the heavily biased Plasmodial pan-genome structure. Specifically, we propose a suitable computational framework for comparative genomics performed by exhaustive all-against-all sequence comparisons, bit-score normalization and clustering while, we validate our methodology using both statistics and biological assessment. As a control group for the assessment of the effect of CBR-masking on the elucidation of the pan-genome structure, the milder (i.e. in terms of CBRs fraction) *Chlamydiales* pan-genome was selected.

Our data indicate that the *Chlamydiales* pan-genome structure is not largely affected by the different CBR-masking approaches being followed. In contrast, the *Plasmodium* pangenome structure is largely dependent on the different strategies used to handle CBRs and we further investigate which are the optimal choices.

**The *Plasmodium* pan-genome**

The completion of MCL runs yielded 20 discrete cluster sets of the pan-genome file (**Table 6**). We analyze the content of the Plasmodia pan-genome in protein families, and seek potential differences depending on how CBRs were handled during the search step. A '*protein family*' is defined as a cluster containing at least 3 proteins and a '*core family*' is a family with at least one protein from each species. We define as '*robust clusters*' those that remain identical in all MCL runs, which corresponds to the most reliably identified families. Finally, we consider the '*unique*' proteins, i.e. those with no similarity within the pan-genome.

In total, the search for robust clusters identified 3186 clusters including 2498 protein families, the remaining 689 clusters corresponding to 569 unique genes and 120 doublets. Remarkably, 1932 robust clusters were among the 'core families', indicating that any estimates on the Plasmodia core pan-genome are largely dependent on handling CBRs (**Table 6**).

**Table 6:** Analysis of the protein families of the *Plasmodium* pan-genome. Columns signified by 1-20 correspond to the discrete MCL runs. Robust clusters (RC) are those clusters that remain invariant between all the runs performed. The OrthoMCL robust (ORC) column correspond to clusters that remain identical both in MCL and OrthoMCL algorithms. Results of Cores, Doublets and unique are provided in the original form (first row of each term) and in percentage (second rom of each term).

| **MCL Results** | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. Pan-genome proteins:** | 50371 | | | | **No. Masked proteins:** | | | 32136 (63.8%) | | | | **No. Robust proteins:** | | | 23605 | | | | **No. Robust proteins (MCL & OrthoMCL): 33330** | | | |
| **MCL** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **RC** | **ORC** |
| **Clusters** | 6780 | 6093 | 6033 | 4904 | 6567 | 6117 | 6148 | 5329 | 6940 | 6322 | 6362 | 5182 | 7036 | 6478 | 6515 | 5415 | 6962 | 6444 | 6383 | 5536 | 3186 | 5367 |
| **Families** | 5080 | 4746 | 4777 | 4115 | 5029 | 4897 | 3894 | 4559 | 5111 | 4886 | 4935 | 4375 | 5172 | 4992 | 5012 | 4587 | 5189 | 5025 | 5016 | 4700 | 2498 | 4033 |
| **#Cores** | 2796 | 2903 | 2919 | 2641 | 3072 | 3308 | 3286 | 3306 | 2744 | 3008 | 2962 | 2824 | 2745 | 3040 | 3013 | 2906 | 2801 | 3042 | 3064 | 3188 | 1932 | 2275 |
| **%Cores** | 55 | 61.2 | 61.1 | 64.2 | 61.1 | 67.5 | 84.4 | 72.5 | 53.7 | 47.6 | 60 | 64.5 | 53.1 | 60.9 | 60.1 | 63.4 | 54 | 60.5 | 61.1 | 67.8 | 75 | 56.4 |
| **#Doublets** | 1069 | 741 | 651 | 207 | 908 | 613 | 649 | 188 | 1198 | 830 | 823 | 225 | 1234 | 878 | 899 | 245 | 1141 | 811 | 762 | 244 | 120 | 747 |
| **%Doublets** | 21 | 15.6 | 13.7 | 5 | 18.1 | 12.5 | 16.7 | 4.1 | 23.4 | 17 | 16.7 | 5.1 | 23.9 | 17.6 | 17.9 | 5.2 | 22 | 16.1 | 15.2 | 5.2 | 6.5 | 18.5 |
| **#Uniques** | 631 | 606 | 605 | 582 | 630 | 607 | 605 | 582 | 631 | 606 | 604 | 582 | 630 | 608 | 604 | 583 | 632 | 608 | 605 | 592 | 569 | 587 |
| **%Uniques** | 12.4 | 12.8 | 12.7 | 14.1 | 12.5 | 12.4 | 15.5 | 12.8 | 12.3 | 9.6 | 12.2 | 13.3 | 12.2 | 12.2 | 12.1 | 17.7 | 12.2 | 12.1 | 12.1 | 12.6 | 23.4 | 14.6 |

In addition, **Figure 12** depicts the robust family size distribution with a peak at low count families (i.e. not all species present) and another peak at the number of genomes compared. The bimodal nature of the distribution follows the shape of similar pan-genome analyses (F. E. Psomopoulos et al., 2012).

The discrete MCL analysis, also, highlights the effects of CBRs in computing the Plasmodium pan-genome. Depending on which filtering mode we followed, we observe different protein families to be formed (**Figure 12**, **Table 6**). For example, in MCL1 (query & database masked) we observe 5080 families to be formed as opposed to MCL4 (query & database unmasked) with only 4904.

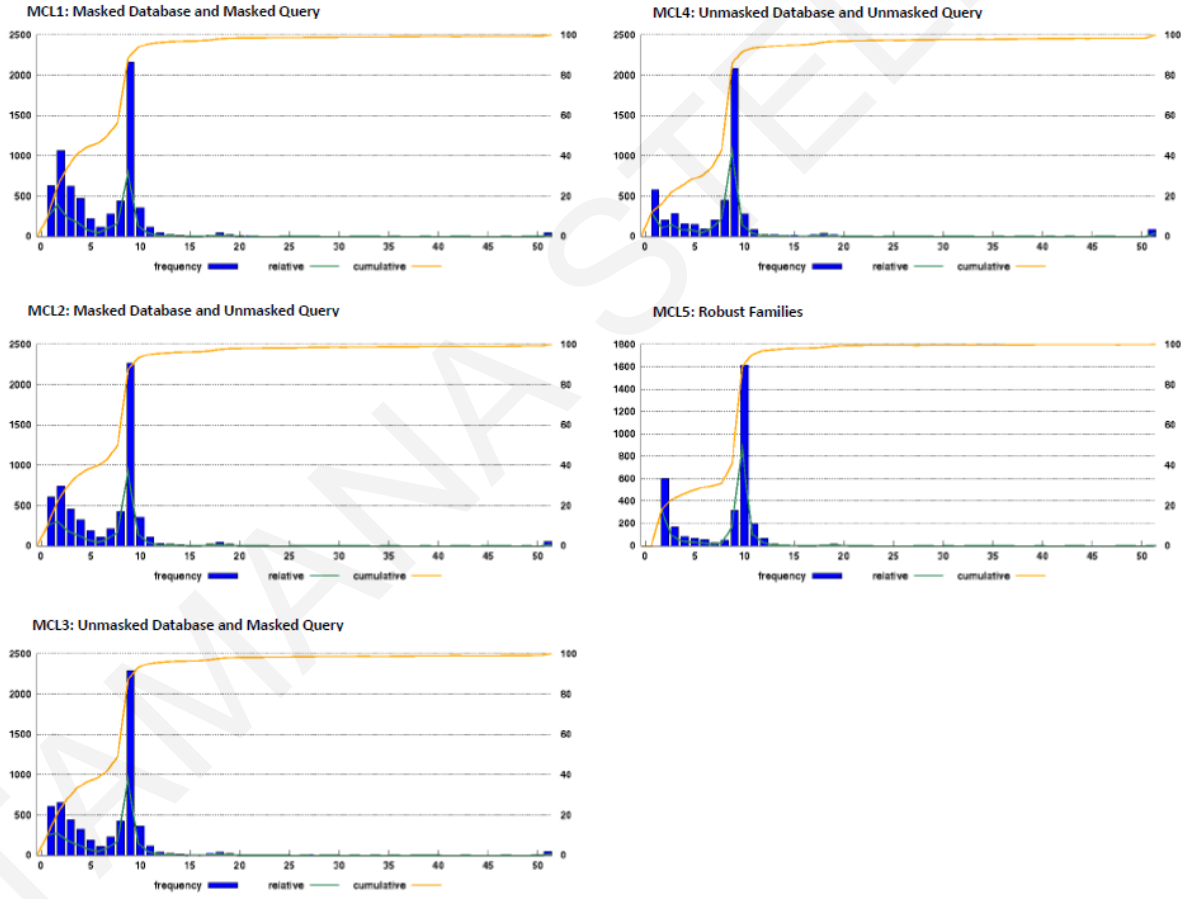



**Figure 12:** Protein family size distribution. Cluster size is displayed on the x-axis (bins until 50 are shown; absolute frequency of clusters is shown on the left y-axis (bars, green curve); cumulative count of clusters is shown on the right y-axis (orange curve). The typical bimodal nature of the distribution can be seen (i.e. one peak at low cluster sizes and one at cluster size approx. equal to the number of genomes analyzed.

The average sequence length distribution of the discrete MCL runs, provide interesting results of the effects of CBRs in computing the *Plasmodium* pan-genome structure. We observe that non-robust clusters are on average composed of longer protein sequences (ranging 1203.50 - 1548.32 amino acids) as opposed to the robust (874.5 amino acids). This suggests that the most reliably identified protein families of the *Plasmodium* pan-genome are composed of average length polypeptides, indicating that longer, multi-domain protein sequences are more difficult to correctly assigned into clusters. (**Table 7**).

**Table 7:** List of average, median and standard error of the protein families as identified by the discrete MCL runs and robust cluster analysis.

| MCL Run | Average | Median | Standard Error | MCL Run | Average | Median | Standard Error |
|---|---|---|---|---|---|---|---|
| **MCL1** | 1479.27 | 1170.5 | 28.91 | **MCL11** | 1472.34 | 1163.5 | 29.80 |
| **MCL2** | 1426.53 | 1127 | 30.28 | **MCL12** | 1257.65 | 996.5 | 29.85 |
| **MCL3** | 1381.21 | 1101.5 | 29.24 | **MCL13** | 1518.75 | 1200.5 | 28.74 |
| **MCL4** | 1203.50 | 946.5 | 31.44 | **MCL14** | 1529.36 | 1193.5 | 30.45 |
| **MCL5** | 1548.32 | 1192 | 31.11 | **MCL15** | 1525.06 | 1202 | 30.23 |
| **MCL6** | 1477.81 | 1128 | 31.22 | **MCL16** | 1330.32 | 1059 | 29.77 |
| **MCL7** | 1507.83 | 1150.5 | 31.82 | **MCL17** | 1516.94 | 1198 | 28.90 |
| **MCL8** | 1322.17 | 1023.5 | 31.70 | **MCL18** | 1520.41 | 1189 | 30.40 |
| **MCL9** | 1514.28 | 1191 | 29.07 | **MCL19** | 1504.57 | 1182.5 | 30.39 |
| **MCL10** | 1504.91 | 1171.5 | 30.87 | **MCL20** | 1368.17 | 1069 | 30.95 |
| **Robust Clusters** | 874.84 | 792 | 15.48 | | | | |

In order to identify which of the MCL runs provide statistically significant results we computed the Wilcoxon Rank Sum test (WRS; (Wilcoxon, 1945). In detail, the WRS test assesses whether two population mean ranks differ.

We observe that with most MCL runs there is a statistically significant difference between them meaning that different protein families are formed. These findings indicate that our perception of the *Plasmodium* family structure largely depends on the CBR-masking approach being followed. In **Table 8**, we list the non-significant results of the WRS test while, the full table of the WRS test results can be found in Appendix I (Table S- 1).

**Table 8:** The results of Wilcoxon Rank Sum test. Only non-significant results are displayed.

| Wilcoxon Rank Sum Test | | | | | |
|---|---|---|---|---|---|
| **MCL vs MCL Run** | | **p-value** | **MCL vs MCL Run** | | **p-value** |
| MCL1 | MCL9 | 0.11 | MCL11 | MCL15 | 0.51 |
| MCL1 | MCL13 | 0.33 | MCL11 | MCL18 | 0.66 |
| MCL1 | MCL17 | 0.33 | MCL11 | MCL19 | 0.19 |
| MCL2 | MCL3 | 0.22 | MCL13 | MCL17 | 0.30 |
| MCL2 | MCL11 | 0.57 | MCL14 | MCL15 | 0.49 |
| MCL2 | MCL19 | 0.47 | MCL14 | MCL19 | 0.20 |
| MCL3 | MCL10 | 0.25 | MCL15 | MCL18 | 0.27 |
| MCL3 | MCL11 | 0.07 | MCL15 | MCL19 | 0.05 |
| MCL3 | MCL14 | 0.07 | MCL18 | MCL19 | 0.38 |
| MCL3 | MCL18 | 0.17 | MCL3 | MCL19 | 0.60 |
| MCL5 | MCL10 | 0.16 | MCL4 | MCL5 | 0.11 |
| MCL5 | MCL11 | 0.45 | MCL5 | MCL14 | 0.42 |
| MCL5 | MCL15 | 0.91 | MCL6 | MCL7 | 0.56 |
| MCL5 | MCL18 | 0.23 | MCL9 | MCL13 | 0.67 |
| MCL9 | MCL17 | 0.54 | MCL10 | MCL19 | 0.52 |
| MCL10 | MCL11 | 0.50 | MCL11 | MCL14 | 0.95 |
| MCL10 | MCL15 | 0.19 | | | |

**Apicoplast analysis**

An earlier study provides evidence that the apicoplast, an organelle present in all *Plasmodium* species, may be the subcellular location of promising drug-candidate genes for the eradication of malaria (Sato et al., 2013; Zuegge et al., 2001). Apicoplasts have been shown to import hundreds of nuclear-encoded proteins (Nisbet and McKenzie, 2016; Waller et al., 2000; Zuegge et al., 2001). Here, we investigate the presence of proteins having a GO cellular component term of apicoplast using the robust clusters.

We found that from the 3186 robust clusters, there are 276 protein families, 232 core and 11 doublets that are composed of proteins with the GO term 'apicoplast'. Furthermore, we divided these clusters into those composed of proteins with experimentally confirmed location and to those with inferred sequence similarity. We observed that, from the 276 and 232 core families, only 34 and 30 families respectively were composed with proteins with experimentally determined location confirming the results of an earlier study (Zuegge et al., 2001).

It is worth mentioning that, none of the unique (clusters composed of only one protein sequence) proteins in our robust dataset were found to have 'apicoplast' as a GO term. This finding suggests that apicoplast proteins are conserved throughout the *Plasmodium* species and provide a clue to the *Plasmodium* evolution. However, a

more thorough investigation of these families is needed. Specifically, regarding their structure and function which could facilitate the identification of more apicoplast proteins and their potential medical importance in the drug/vaccine development pipeline.

***MCL versus OrthoMCL Results***

In total, the search for robust MCL-OrthoMCL clusters identified 5367 robust clusters, where 4033 are protein families (**Table 6**). Out of those families 747 correspond to doublets and 587 are unique genes. Interestingly, we observe 2275 (i.e. 56.4% of the families) robust clusters among the 'core families', indicating that these families have an important role at the parasite life cycle and share a common evolutionary history. A comparative genomic analysis of six of the genomes used in our study and similar methodology (except *P. cynomolgi* B and *P. falciparum* IT) revealed that the core genome is comprised of 3,351 (corresponding to 22%-65% of each genome) orthologous genes (Cai et al., 2010). The difference between our results to their results could be due to the slightly different approach we follow and the additional genomes in our dataset. Cai and colleagues identified the core orthologous genes using OrthoMCL only, while we report the robust core genes of both MCL and OrthoMCL.

Following the approach described in (Carlton, 2006; Hall et al., 2005), we created a list with the robust MCL-OrthoMCL clusters (**Table 9**). Remarkably, we observe differences when comparing the families reported in (Carlton, 2006) with this work. An example is the yir/cir/bir family, which is a large variant gene family predicted to be involved in antigenic variation (Carlton et al., 2002; del Portillo et al., 2001). In the original list, compiled by four out of the nine complete sequenced *Plasmodium* species, this family shows various copies only in *P. chabaudi*, *P. yoelii* and *P. berghei*. In our case, we observe all species in the pan-genome to have at least one such protein homolog. Another example is the case of the rhoptry protein (reticulocyte binding protein) family, where we observe increased copies in all *Plasmodium* species of our dataset. Rhoptry proteins contribute to host cell recognition and junction formation by the merozoite (Bapat et al., 2011) and the fact we observed multiple copies preserved throughout the malaria parasite genomes highlights their

biological importance in *Plasmodium* signal transduction pathways during erythrocyte invasion (Gunalan et al., 2013).

Protein families found to be absent from this list are the pyst-a/b/c, pcst-f/g/h, HAD-Hydrolase and pst-a. This discrepancy is possibly due to the fact that these families contain members highly variable in lengths and composition bias content, thus they are clustered differently with different strategies.

Additionally, we compiled a complete list of protein description for all MCL-OrthoMCL robust genes, but more work is needed for further analysis.

**Table 9:** *Plasmodium* gene families (adapted from (Carlton, 2006)). Comparison of robust genes found from MCL & OrthoMcl.

| Family Name | P. berghei ANKA | P. chabaudi Chabaudi | P. cynomolgi B | P. falciparum 3D7 | P. falciparum IT | P. knowlesi H | P. vivax Sal1 | P. yoelii 17XNL | P. yoelii YM | Function |
|---|---|---|---|---|---|---|---|---|---|---|
| **Pf-fam-b** | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | Phist protein |
| **Pf-fam-e** | 0 | 0 | 3 | 0 | 0 | 2 | 3 | 0 | 0 | RAD protein |
| **Pf-fam-i** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Acyl-Co- A synthetase |
| **Var** | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | Erythrocyte membrane antigen (PfEMP1) |
| **etramp** | 2 | 2 | 7 | 6 | 6 | 7 | 7 | 2 | 2 | Early transcribed membrane protein |
| **rhoph/clag** | 2 | 3 | 2 | 5 | 6 | 2 | 3 | 3 | 2 | Cytoadherence linked asexual protein |
| **Rhoptry protein** | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 10 | 7 | Reticulocyte binding protein |
| **Rifin/stevor** | 0 | 0 | 0 | 185 | 138 | 0 | 0 | 1 | 1 | Variant RBC surface antigen |
| **Yir/cir/bir/kir** | 2 | 2 | 5 | 1 | 1 | 1 | 1 | 2 | 2 | Variant antigen |

***Validation***

**Protein domain architectures**

So far, our analysis was based on the 20 different possible structures of the *Plasmodium* pan-genome as provided by our computational pipeline. However, we need a biological approach that could pinpoint the best CBR filtering for comparative analysis of compositionally biased pan-genomes. A possible measure for rating and identifying the best approach is by using protein domain architectures. For this purpose, we choose the Pfam database (version 27.0; (Finn et al., 2016, 2006)), which is a collection of protein domain families.

The results of each MCL run were fed into PfamScan tool (Finn et al., 2006) with default parameters that retrieve for each cluster the Pfam protein domains. In total, we identified 2020 unique Pfam protein domains out of 47891 Pfam domains.

Following the identification of the Pfam domains we check for the homogeneity of domain architectures within each cluster by computing a Domain Composition Homogeneity score (DCH-score, ranging between 0-100) for each cluster based on how many different domains were found and the number of proteins having these domains. The DCH score is computed as follows:

$$DCH = \frac{100}{n}\sum_{i=1}^{n}\frac{m_i}{m} \quad (\mathbf{4}),$$ where:

*n* is the number of sequences in a cluster, *m* is the number of distinct Pfam domain types in the cluster and $m_i$ is the corresponding number of Pfam domains in the i-th sequence. Clearly, this approach does not take into account the domain order (architecture) or possible domain fusion events, but provides a fast, intuitive proxy for validating the quality of detected clusters. An example depicting the computation of DCH-score can be found in Figure 13.

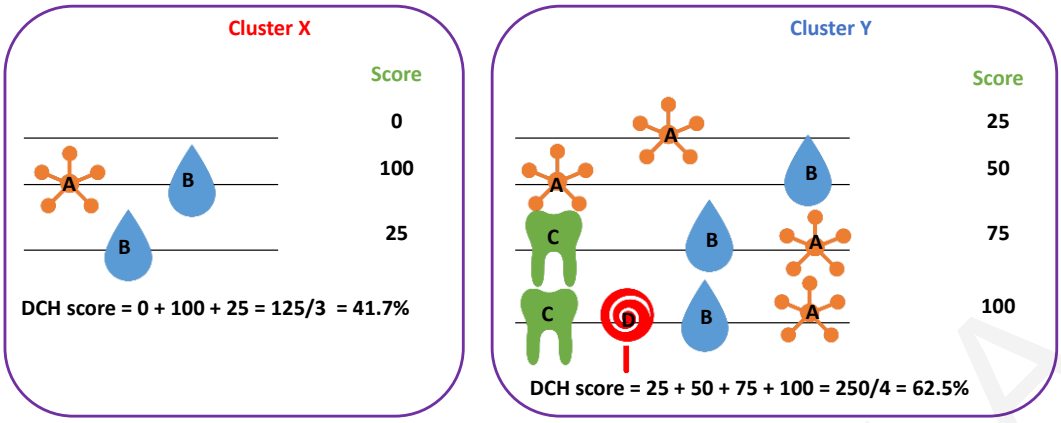



**Figure 13:** An example of how the Domain Composition Homogeneity (DCH) score is calculated. Obviously, cases of domain fusion events may lower the computed DCH score for a particular family.

Furthermore, the computations were divided into two sets of scenarios. In the first set of calculations, we included all clusters of each run and the scoring scheme was set to 100 for clusters for which no domains were detected. In the second scenario, we excluded all singleton clusters (NS) and clusters where no domains detected (NSD) (**Table 10**). Following the calculations for each cluster of the discrete MCL runs we rated each run by calculating the average, standard deviation, standard error scores (**Table 11**).

**Table 10:** The number of clusters of each MCL run counted at each step of the protein domains architecture analysis. **#Clusters:** number of all clusters; **#NS:** number of clusters after excluding all singleton clusters; **#NSD:** number of clusters after excluding both singleton clusters and clusters with no domains.

| MCL Run | #Clusters | #NS | #NSD | MCL Run | #Clusters | #NS | #NSD |
|---|---|---|---|---|---|---|---|
| **MCL1** | 6780 | 6149 | 3336 | **MCL11** | 6362 | 5758 | 3185 |
| **MCL2** | 6093 | 5487 | 3063 | **MCL12** | 5182 | 4600 | 2738 |
| **MCL3** | 6033 | 5428 | 3066 | **MCL13** | 7036 | 6406 | 3494 |
| **MCL4** | 4904 | 4322 | 2608 | **MCL14** | 6478 | 5870 | 3255 |
| **MCL5** | 6567 | 5937 | 3187 | **MCL15** | 6515 | 5911 | 3258 |
| **MCL6** | 6117 | 5510 | 3061 | **MCL16** | 5415 | 4832 | 2852 |
| **MCL7** | 6148 | 5543 | 3069 | **MCL17** | 6962 | 6330 | 3448 |
| **MCL8** | 5329 | 4747 | 2866 | **MCL18** | 6444 | 5836 | 3233 |
| **MCL9** | 6940 | 6309 | 3426 | **MCL19** | 6383 | 5778 | 3206 |
| **MCL10** | 6322 | 5716 | 3162 | **MCL20** | 5536 | 4944 | 2910 |

We ranked the results by the average DCH-score, ranking as 1 the worst performing strategy and as 20 the best performing one. Remarkably, both of the strategies ranked as the best performing are those that the sequence comparisons plus r-fraction steps were employed with the masked sequences. In particular, in the “all clusters” case, we observe MCL9 (bl2seq: masked database/query; average DCH score: 93.46) to be the best performing strategy while, in the “no singletons and no domains” case, the higher average of the DCH score is 86.79 (MCL13; CSC: masked database/query). Since the number of unique proteins detected among the different MCL schemes does not vary too much (582 - 632) we expect that it is the families with no domains that boost the performance of MCL9 which could be the optimal strategy for identifying the *Plasmodium* pan-genome unique proteins. However, MCL13 is compares well in respect to performance quality. Furthermore, as the worst performing strategy was ranked the MCL20 (CSC: unmasked database/query; average DCH score: 91.36; CBS: unmasked database/query; average DCH score: 83.56) indicating that heavily biased pan-genomes require more careful CBR-filtering.

**Table 11:** Ranking of the average DCH scores for each MCL run. **All:** all clusters were included in the calculations; **NSD:** singletons and clusters with no domains were excluded. The Rank column signifies the order of the lower up to the higher average DCH score. Green: best performing strategies; Purple: worst performing strategies.

| MCL Run | All | | NSD | | Standard Deviation | | Standard Error | |
|---|---|---|---|---|---|---|---|---|
| | Average | Rank | Average | Rank | All | NSD | All | NSD |
| MCL1 | 93.31 | 17 | 86.41 | 17 | 17.76 | 23.39 | 0.22 | 0.28 |
| MCL2 | 92.67 | 9 | 85.43 | 8 | 18.82 | 24.48 | 0.24 | 0.31 |
| MCL3 | 92.63 | 7 | 85.50 | 9 | 18.78 | 24.31 | 0.24 | 0.31 |
| **MCL4** | **91.45** | **2** | **83.96** | **2** | **20.84** | **26.39** | **0.28** | **0.38** |
| MCL5 | 93.19 | 16 | 85.96 | 14 | 17.78 | 23.45 | 0.22 | 0.29 |
| MCL6 | 92.59 | 6 | 85.19 | 6 | 18.51 | 23.98 | 0.24 | 0.31 |
| MCL7 | 92.63 | 8 | 85.24 | 7 | 18.47 | 23.96 | 0.24 | 0.31 |
| MCL8 | 91.79 | 3 | 84.74 | 4 | 19.55 | 24.55 | 0.28 | 0.34 |
| **MCL9** | **93.46** | **20** | **86.75** | **19** | **17.49** | **23.04** | **0.21** | **0.28** |
| MCL10 | 92.88 | 12 | 85.77 | 11 | 18.25 | 23.77 | 0.23 | 0.30 |
| MCL11 | 92.99 | 14 | 86.00 | 15 | 18.03 | 23.49 | 0.23 | 0.29 |
| MCL12 | 91.85 | 4 | 84.57 | 3 | 19.79 | 25.07 | 0.27 | 0.35 |
| **MCL13** | **93.44** | **19** | **86.79** | **20** | **17.49** | **22.98** | **0.21** | **0.27** |
| MCL14 | 92.92 | 13 | 85.92 | 13 | 18.14 | 23.58 | 0.23 | 0.29 |
| MCL15 | 93.03 | 15 | 86.07 | 16 | 17.97 | 23.42 | 0.23 | 0.29 |
| MCL16 | 91.97 | 5 | 84.76 | 5 | 19.46 | 24.69 | 0.26 | 0.34 |

| MCL Run | All | | NSD | | Standard Deviation | | Standard Error | |
|---|---|---|---|---|---|---|---|---|
| | Average | Rank | Average | Rank | All | NSD | All | NSD |
| MCL17 | 93.37 | 18 | 86.61 | 18 | 17.61 | 23.14 | 0.21 | 0.28 |
| MCL18 | 92.87 | 11 | 85.80 | 12 | 18.25 | 23.73 | 0.28 | 0.30 |
| MCL19 | 92.81 | 10 | 85.68 | 10 | 18.35 | 23.84 | 0.23 | 0.30 |
| **MCL20** | **91.36** | **1** | **83.56** | **1** | **20.09** | **25.30** | **0.27** | **0.34** |

Following the statistical analysis, we calculate the WRS test using the average values from the previous step (Table S-2). Regarding the best performing strategies (MCL9/MCL13) depicted by the ranking of the average DCH-scores, we observed that WRS test shows statistically significant difference with MCL4/6/7/8/12/16/20 for both runs. Thus, both modes are creating statistically significant different instances of the *Plasmodium* pan-genome structure from MCL modes where sequence comparisons and/or R-fraction normalization was performed with unmasked sequences (i.e. db and/or query sequence was unmasked). In contrast, the worst performing strategy based on the ranking of average DCH scores, MCL20 (CSC: unmasked database/query), the WRS test showed statistically significant difference to all other runs except MCL4/8/12/16 (i.e. db and/or query sequence was unmasked). Therefore, these results demonstrate that extremely biased pan-genomes, such as malaria parasites, require more careful CBR-filtering and this could be achieved when the sequence comparisons plus r-fraction normalization are computed with the masked sequences prior the clustering step.

**Major *Plasmodium* protein Families**

Independently of the protein domain architectures, we further validate the *in silico* comparative genomics pipeline with the Major *Plasmodium* protein Families (MPFs), such as the rifin/stevor, yir/cir/bir and PfEMP1 (var). These are large protein families that are well-known for their medical importance as vaccine candidates but also, they are extremely high in CBRs (Carlton, 2006).

We analyzed the discrete MCL runs by computationally extracting the functional annotations focusing only on the MPFs. Specifically, we extracted only the proteins that were annotated as rifin, stevor, surfin, yir/cir/bir, PfEMP1 (var), SICAvar or Pfmc-2TM from a list consisting of all the proteins in our pan-genome file. In details, an in-house developed script takes as input a list of all proteins in our pan-genome dataset along with their functional annotation and extracts the protein ids and annotation focusing on the MPFs only. Then, for each MCL run and for each MPF the script identifies all clusters containing at least one member of the family. Finally, for each run the script outputs a file with all the clusters consisting with that MPF and a summary file displaying the number of clusters found, the number of proteins in each cluster and the number of species that are present in each run.

The analysis was statistically verified using the Adjusted Rand Index (ARI; (Hubert and Arabie, 1985; Steinley, 2004)). Briefly, the ARI is a form of the Rand Index and can be

defined as an adjusted measure for the chance-grouping of elements between two clusterings. For example, given the set of all the proteins in our dataset annotated as rifins, and two clusterings of these proteins (e.g. MCL1 and MCL2), we first computed a contingency table that summarizes the number of rifins in common between MCL1 and MCL2 and then calculated ARI.

The ARI analysis showed that for the rifin and Pfmc-2TM family all strategies perform equally well (data not shown). The ARI value of these families for all MCL runs was one indicating that the data clusters are exactly the same between two clustering modes. For the PfEMP1, SICAvar, Stevor and yir/cir/bir protein families the ARI values ranging from 0.83-1, indicating that most clusters are similar between two clusterings. MCL9 (bl2seq: masked database/query) and MCL13 (CSC: masked database/query) which, have the highest average scores in the Pfam analysis, also, appear to be the best performing strategies for handling CBRs. In **Figure 14**, we provide a heat map of the ARI values of the PfEMP1 family as a representative example for all families depicting high ARI values.

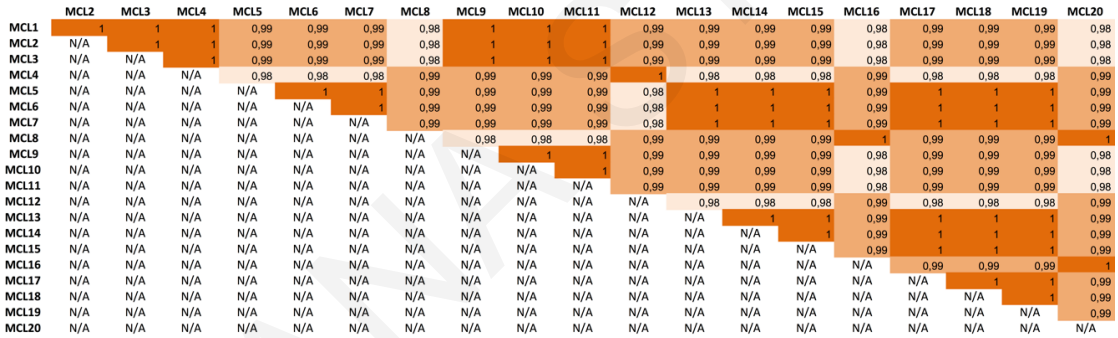



**Figure 14:** A heat map of the Adjusted Rand Index values of the PfEMP1 family.

Remarkably, the *Surfin* family had the most variable ARI values ranging from 0.48-1 (**Figure 15**) indicating that different instances of this family are created between two clustering's. We observed that 2 clusters are formed between the discrete MCL runs, one that groups modes creating similar instances of the *Surfin* family and another group displaying different instances of the surfing family between two MCL runs. Specifically, in the cluster grouping the MCL13-20 modes we observed that, nearly half of the clusters are composed of different *Surfin* proteins as opposed to the remaining MCL modes. The differences between them lies in that MCL13-20 were computed using the R-fraction normalization employed with the self-comparison mode instead of the bl2seq mode. Thus, highlighting how we compute the R-fraction normalization might, also, affect the robustness of comparative genomics. *Surfins* is a small gene family encoded by a family of

ten surf genes, including three pseudo genes, implicated in the invasion of erythrocytes by merozoites but also, extremely high in CBRs (Winter et al., 2005).

| | MCL2 | MCL3 | MCL4 | MCL5 | MCL6 | MCL7 | MCL8 | MCL9 | MCL10 | MCL11 | MCL12 | MCL13 | MCL14 | MCL15 | MCL16 | MCL17 | MCL18 | MCL19 | MCL20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCL1 | 1 | 1 | 1 | 0,7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,48 | 0,48 | 0,48 | 0,76 | 0,48 | 0,48 | 0,54 | 0,76 |
| MCL2 | N/A | 1 | 1 | 0,7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,48 | 0,48 | 0,48 | 0,76 | 0,48 | 0,48 | 0,54 | 0,76 |
| MCL3 | N/A | N/A | 1 | 0,7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,48 | 0,48 | 0,48 | 0,76 | 0,48 | 0,48 | 0,54 | 0,76 |
| MCL4 | N/A | N/A | N/A | 0,7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,48 | 0,48 | 0,48 | 0,76 | 0,48 | 0,48 | 0,54 | 0,76 |
| MCL5 | N/A | N/A | N/A | N/A | 0,7 | 0,7 | 0,69 | 0,7 | 0,7 | 0,7 | 0,7 | 0,57 | 0,57 | 0,57 | 0,5 | 0,57 | 0,57 | 0,54 | 0,5 |
| MCL6 | N/A | N/A | N/A | N/A | N/A | 1 | 1 | 1 | 1 | 1 | 1 | 0,49 | 0,49 | 0,49 | 0,76 | 0,49 | 0,49 | 0,56 | 0,76 |
| MCL7 | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 1 | 1 | 1 | 1 | 0,49 | 0,49 | 0,49 | 0,76 | 0,49 | 0,49 | 0,56 | 0,76 |
| MCL8 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 1 | 1 | 1 | 0,49 | 0,49 | 0,49 | 0,76 | 0,49 | 0,49 | 0,56 | 0,76 |
| MCL9 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 1 | 1 | 0,49 | 0,49 | 0,49 | 0,76 | 0,49 | 0,49 | 0,56 | 0,76 |
| MCL10 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 1 | 0,49 | 0,49 | 0,49 | 0,76 | 0,49 | 0,49 | 0,56 | 0,76 |
| MCL11 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 0,48 | 0,48 | 0,48 | 0,76 | 0,48 | 0,48 | 0,54 | 0,76 |
| MCL12 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,49 | 0,49 | 0,49 | 0,76 | 0,49 | 0,49 | 0,56 | 0,76 |
| MCL13 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 1 | 0,66 | 1 | 1 | 0,92 | 0,66 |
| MCL14 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 0,66 | 1 | 1 | 0,92 | 0,66 |
| MCL15 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,66 | 1 | 1 | 0,92 | 0,66 |
| MCL16 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,66 | 0,66 | 0,74 | 1 |
| MCL17 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1 | 0,92 | 0,66 |
| MCL18 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,92 | 0,66 |
| MCL19 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,74 |
| MCL20 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

**Figure 15:** A heat map of the Adjusted Rand Index values for the Surfin family.

In addition to the MPFs ARI validation analysis, we computed the ARI for all clusters of the discrete MCL runs (**Figure 16**). Here, the ARI values ranging from 0.4-0.9 strongly suggest that different protein families are formed in each run and further support our observation that any estimates on the *Plasmodia* pan-genome structure are largely dependent on how we handle CBRs. The best strategies that appear to perform equally well are the MCL9 (bl2seq: masked database/query) and MCL13 (CSC: masked database/query).

| | MCL2 | MCL3 | MCL4 | MCL5 | MCL6 | MCL7 | MCL8 | MCL9 | MCL10 | MCL11 | MCL12 | MCL13 | MCL14 | MCL15 | MCL16 | MCL17 | MCL18 | MCL19 | MCL20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCL1 | 0,9 | 0,84 | 0,64 | 0,84 | 0,85 | 0,85 | 0,57 | 0,97 | 0,92 | 0,9 | 0,51 | 0,88 | 0,88 | 0,88 | 0,4 | 0,88 | 0,88 | 0,88 | 0,66 |
| MCL2 | N/A | 0,8 | 0,61 | 0,79 | 0,79 | 0,79 | 0,55 | 0,89 | 0,91 | 0,84 | 0,49 | 0,81 | 0,82 | 0,83 | 0,39 | 0,82 | 0,83 | 0,83 | 0,63 |
| MCL3 | N/A | N/A | 0,63 | 0,76 | 0,77 | 0,77 | 0,54 | 0,81 | 0,8 | 0,86 | 0,49 | 0,78 | 0,79 | 0,8 | 0,39 | 0,79 | 0,79 | 0,8 | 0,62 |
| MCL4 | N/A | N/A | N/A | 0,58 | 0,58 | 0,58 | 0,47 | 0,63 | 0,62 | 0,64 | 0,52 | 0,59 | 0,6 | 0,6 | 0,39 | 0,59 | 0,6 | 0,6 | 0,52 |
| MCL5 | N/A | N/A | N/A | N/A | 0,99 | 0,99 | 0,63 | 0,86 | 0,83 | 0,84 | 0,46 | 0,94 | 0,94 | 0,94 | 0,41 | 0,94 | 0,94 | 0,94 | 0,69 |
| MCL6 | N/A | N/A | N/A | N/A | N/A | 0,99 | 0,64 | 0,86 | 0,84 | 0,85 | 0,47 | 0,94 | 0,95 | 0,94 | 0,42 | 0,94 | 0,95 | 0,94 | 0,7 |
| MCL7 | N/A | N/A | N/A | N/A | N/A | N/A | 0,64 | 0,86 | 0,84 | 0,85 | 0,47 | 0,94 | 0,94 | 0,94 | 0,41 | 0,94 | 0,94 | 0,94 | 0,7 |
| MCL8 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,57 | 0,56 | 0,57 | 0,43 | 0,6 | 0,61 | 0,62 | 0,5 | 0,61 | 0,61 | 0,62 | 0,65 |
| MCL9 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,94 | 0,91 | 0,51 | 0,9 | 0,89 | 0,9 | 0,4 | 0,9 | 0,89 | 0,9 | 0,67 |
| MCL10 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,88 | 0,5 | 0,87 | 0,88 | 0,88 | 0,4 | 0,86 | 0,87 | 0,88 | 0,66 |
| MCL11 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,51 | 0,88 | 0,88 | 0,89 | 0,4 | 0,87 | 0,88 | 0,89 | 0,67 |
| MCL12 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,48 | 0,48 | 0,48 | 0,41 | 0,48 | 0,48 | 0,48 | 0,47 |
| MCL13 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,98 | 0,98 | 0,42 | 0,99 | 0,98 | 0,97 | 0,72 |
| MCL14 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,98 | 0,43 | 0,98 | 0,99 | 0,98 | 0,72 |
| MCL15 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,43 | 0,97 | 0,98 | 0,99 | 0,73 |
| MCL16 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,42 | 0,43 | 0,43 | 0,54 |
| MCL17 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,98 | 0,98 | 0,72 |
| MCL18 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,98 | 0,72 |
| MCL19 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0,73 |
| MCL20 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

**Figure 16:** A heat map of the Adjusted Rand Index values using all clusters from the discrete MCL runs.

Combining all statistical analyses (length distribution, average DCH scores, WRS test) along with the ARI results we observe that the consistently best performing strategies for computing extremely biased pan-genomes are MCL9 (bl2seq: masked database/query) and MCL13 (CSC: masked database/query). Thus, we propose that the best strategy when dealing with heavily biased pan-genomes is the sequence comparisons plus R-fraction normalization to be employed with the masked sequences prior the clustering step.

### *Plasmodium* vs. *Chlamydiales* pan-genome Analysis

In addition to the extremely compositionally biased pan-genome of *Plasmodium* we examine our methodology using the milder (i.e. in terms of CBRs fraction) *Chlamydiales*

pan-genome as a control group. Most members of the order *Chlamydiales* are obligate intracellular bacteria and important pathogens of humans and animals (Fotis E. Psomopoulos et al., 2012; Wyrick, 2000). Their genome has a smaller fraction of CBRs (approximately 16%; Table 12) and the wide variety of chlamydial genome sizes makes it the ideal taxon for the development of pan-genome analysis methods (Angiuoli et al., 2011; F. E. Psomopoulos et al., 2012). The *Chlamydiales* pan-genome dataset was retrieved from an earlier study (Fotis E. Psomopoulos et al., 2012) and the comparative genomic analysis was performed using our computational method and tools (**Figure 6**).

The respective analysis for the *Chlamydiales* pan-genome revealed very interesting results as we observed very small fraction of differences between the discrete MCL runs comparing to the *Plasmodium* pan-genome analysis (Table 12). The most important one is that in the milder biased *Chlamydiales* pangenome, 5681 (93%) — 5774 (94.6%) of the clusters were robust. This means that no matter how CBRs filtering was performed in the *Chlamydiales* pan-genome, few clusters were different. In contrast, in *Plasmodium* pan-genome analysis the number of clusters that were among the robust, ranges from 5415 (59.9%) to 7036 (45.3%). Thus, proving that, the robustness of comparative genomics is largely depending on the CBR filtering approach being followed.

This observation is further supported by the descriptive statistics of the average scores of the protein domain architecture analysis (Table 13). Here, we observe that the differences of the average scores are very small between each run in either scenario we tested as opposed to the *Plasmodium* results. In addition, the results of WRS test for the protein domain architecture are *Non-significant* for all runs meaning that all CBR filtering modes perform equally well. Suggesting that, masking CBRs prior the sequence comparison step should be sufficient for the comparative genomics analysis if we have mild compositionally biased genomes.

**Table 12:** Analysis of the families of the *Chlamydiales* pangenome. Columns signified by 1-20 correspond to the discrete MCL runs. Robust clusters (RC) are those clusters that remain invariant between all the runs performed.

| MCL Results | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. Pangenome proteins: 43.736** | | | | | | | **No. Masked proteins: 6.906 (15.8%)** | | | | | | | **No. Robust proteins: 38414 (87.3%)** | | | | | | | |
| **MCL** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **RC** |
| **#Clusters** | 5713 | 5676 | 5681 | 5686 | 5774 | 5726 | 5733 | 5737 | 5767 | 5730 | 5730 | 5739 | 5775 | 5728 | 5737 | 5740 | 5770 | 5732 | 5732 | 5741 | 5373 |
| **#Families** | 2204 | 2192 | 2194 | 2198 | 2227 | 2206 | 2211 | 2216 | 2222 | 2210 | 2210 | 2214 | 2224 | 2205 | 2210 | 2215 | 2222 | 2209 | 2209 | 2216 | 1937 |
| **#Core** | 465 | 467 | 466 | 465 | 462 | 465 | 464 | 465 | 462 | 465 | 464 | 464 | 462 | 465 | 465 | 464 | 462 | 465 | 464 | 463 | 435 |
| **#Doublets** | 1354 | 1345 | 1346 | 1345 | 1392 | 1381 | 1381 | 1377 | 1390 | 1381 | 1379 | 1381 | 1396 | 1384 | 1386 | 1381 | 1393 | 1384 | 1382 | 1384 | 1302 |
| **#Singletons** | 2155 | 2139 | 2141 | 2143 | 2155 | 2139 | 2141 | 2144 | 2155 | 2139 | 2141 | 2144 | 2155 | 2139 | 2141 | 2144 | 2155 | 2139 | 2141 | 2144 | 2134 |

**Table 13:** List of the average, standard deviation and standard error DCH-scores of each MCL run. The split column of each statistical measure corresponds to the two sets of scenarios we tested for the protein domains architecture.

| MCL Run | Average | | Standard Deviation | | Standard Error | |
|---|---|---|---|---|---|---|
| | All | NSD | All | NSD | All | NSD |
| **MCL1** | 96.68 | 90.52 | 12.76 | 20.15 | 0.17 | 0.45 |
| **MCL2** | 96.64 | 90.43 | 12.84 | 20.24 | 0.17 | 0.45 |
| **MCL3** | 96.65 | 90.46 | 12.82 | 20.22 | 0.17 | 0.45 |
| **MCL4** | 96.62 | 90.36 | 12.91 | 20.38 | 0.17 | 0.46 |
| **MCL5** | 96.56 | 90.27 | 12.99 | 20.41 | 0.17 | 0.45 |
| **MCL6** | 96.54 | 90.22 | 13.04 | 20.47 | 0.17 | 0.45 |
| **MCL7** | 96.54 | 90.23 | 13.03 | 20.46 | 0.17 | 0.45 |
| **MCL8** | 96.55 | 90.24 | 13.02 | 20.45 | 0.17 | 0.45 |
| **MCL9** | 96.57 | 90.27 | 12.97 | 20.39 | 0.17 | 0.45 |
| **MCL10** | 96.55 | 90.23 | 13.01 | 20.42 | 0.17 | 0.45 |
| **MCL12** | 96.55 | 90.23 | 13.01 | 20.43 | 0.17 | 0.45 |
| **MCL12** | 96.55 | 90.23 | 13.03 | 20.47 | 0.17 | 0.45 |
| **MCL13** | 96.56 | 90.27 | 13.00 | 20.43 | 0.17 | 0.45 |
| **MCL14** | 96.54 | 90.22 | 13.05 | 20.49 | 0.17 | 0.46 |
| **MCL15** | 96.55 | 90.24 | 13.03 | 20.47 | 0.17 | 0.45 |
| **MCL16** | 96.55 | 90.23 | 13.04 | 20.49 | 0.17 | 0.46 |
| **MCL17** | 96.57 | 90.28 | 12.98 | 20.40 | 0.17 | 0.45 |
| **MCL18** | 96.54 | 90.23 | 13.02 | 20.45 | 0.17 | 0.45 |
| **MCL19** | 96.54 | 90.22 | 13.03 | 20.46 | 0.17 | 0.45 |
| **MCL20** | 96.55 | 90.23 | 13.04 | 20.48 | 0.17 | 0.45 |

**Concluding remarks**

Summing up, in this section we demonstrate that handling CBRs prior the sequence comparison step is an important step in pan-genome analysis, especially in the cases of heavily biased datasets, e.g. *Plasmodium.* We validated several strategies with respect to handling CBRs, concluding that masking both the query and the database sequences is essential for obtaining high-quality clusters. We also highlight the importance of bit-score normalization as a key step in the clustering procedure, at least when the MCL algorithm is employed for this task.

## 3.2. Unique genes in malaria parasites: a pan-genomic approach

In the previous section, we proposed as one of the optimal *in silico* strategies for comparative genomics, the MCL9 (Masked DB/Query (BLASTP) + R-fraction (bl2seq: Masked DB/Query)). We subsequently employ the acquired knowledge on the elucidation of the *Plasmodium* pan-genome unique proteins. The unique proteins of the *Plasmodium* pan-genome are of utmost interest in efforts to explore and understand the molecular biology of malaria parasites. Importantly, this analysis was performed in an updated set of nineteen sequenced *Plasmodium* genomes (see Data and Methods; **Table 2**). We follow the definition of species and strains for *Plasmodium* based on genetic terms as described by (Carlton, 2006; Perkins, 2000).

For clarity, we define as *orphan/strain specific* proteins those that are found to exist in only one strain of *Plasmodium*, e.g. *P. falciparum* 3D7, as *species specific* those that seem to be in only one *Plasmodium* species, e.g. *P. falciparum* and as *Taxonomically Restricted Genes (TRG)* those that seem to be present only in closely-related *Plasmodium* species as defined in (Khalturin et al., 2009).

**The *Plasmodium* pan-genome unique proteins**

The MCL clustering (using the default parameters) analysis of the *Plasmodium* pan-genome yielded 1201 unique proteins, while the sequence comparison step against the NR database found statistically significant hits for 1192 proteins (Figure 17). The remaining 7 did not have any BLASTP/tBLASTN hit and the other 2 proteins did not have any BLASTP hit but had statistically significant hits against the NT database (Table 14).

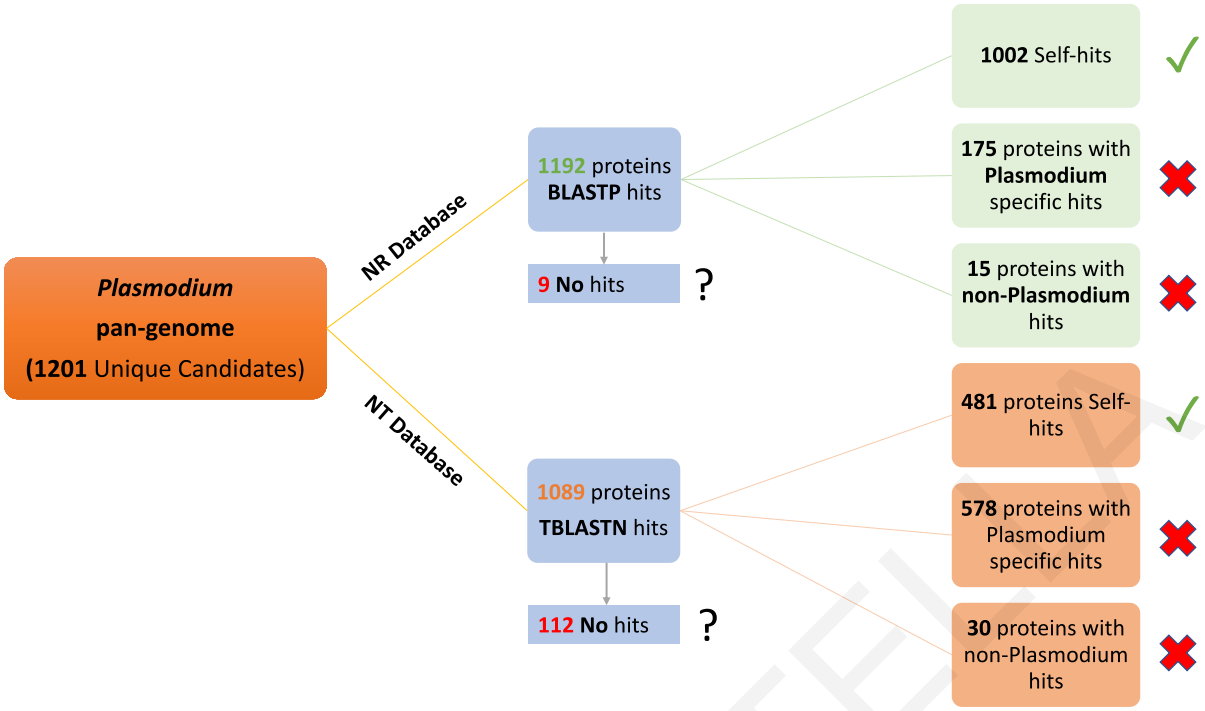



**Figure 17:** Flowchart of the methodology followed for the analysis of the *Plasmodium* pan-genome unique proteins. The check mark symbol denoted those proteins that remained unique in all sequence comparison steps we performed, while, the cross symbol denotes those proteins that are not unique after the NR/NT sequence database comparisons and the **question mark** symbol denotes ambiguous results.

One out of the nine proteins with no hit in the NR database, was annotated as PfEMP1 pseudogene, while the rest are annotated as hypothetical proteins. The PfEMP1 is one of the well-known *Major Plasmodium Protein Families* (MPFs; (del Portillo et al., 2001; Fougère et al., 2017; Frech and Chen, 2011; Malcolm J. Gardner et al., 2002) and it is interesting the fact that the clustering analysis set this protein as unique instead with its respective MPF. Consequently, the next step in our analysis was to examine why these proteins did not cluster along with other proteins (i.e. verify the unique status) or its respective MPF and verify their status. Using the tBLASTN tool from the standalone BLAST suit of tools, we searched if the genes encoding these proteins are present in other *Plasmodia* (or other species) and have not yet been annotated as genes or perhaps are present but lost their function (i.e. verify that are pseudogenes).

*Plasmodium falciparum* IT genome is not deposited in NCBIs' database (Genome and National Center for Biotechnology Information, 2018) and thus, we could not exclusively rely on NCBIs' BLASTP/tBLASTN results. For this reason, we searched PlasmoDB annotated proteins and genomes using the integrated BLAST tools with default settings (e-value: 10,

filtering off). This high e-value threshold permits identification of marginal sequence similarities that are further validated manually, to rule out the possibility that we are losing partially predicted proteins (fragments) or cases of incorrect gene predictions e.g. frameshifts. First, we searched PlasmoDB's annotated proteins datasets using both the unmasked and masked protein sequence of this *P. falciparum* protein. Our initial BLASTP search (for both masked and unmasked sequences) verified that this protein is unique inside *Plasmodium* pan-genome and confirmed our pan-genome results. Then, we performed tBLASTN search against PlasmoDBs genomes datasets (using the unmasked protein sequence) as to check if the gene is present in other *Plasmodia*. This step confirmed the functional annotation of 'PfEMP1 pseudogene' as the top hits from *P. falciparum* 3D7 were annotated as 'PfEMP1 pseudogene'. To cross-validate these results, we performed BLASTX search (compares a nucleotide sequence against a protein database) using its genomic DNA sequence. Inspecting the BLASTX results we noticed frameshift disruptions as the self-hit alignment was in +1 frame and with *P. falciparum* 3D7 this sequence is translated in +2 frame. Combining these results with the scoring criteria computed for this protein, where it has high %CBR, low %GC and is found in a least conserved neighborhood (ND = 0; Table 14) we concluded that this protein sequence is the result of a gene prediction artefact and was excluded from further analysis.

From the remaining eight proteins (all annotated as "hypothetical protein"), we could verify the status of one as strain specific, while the remaining seven are partial sequences (**Table 14**). Regarding the *P. vivax* Sal-1 protein, we confirmed its strain specific status through PlasmoDB's tools (BLASTP, tBLASTN and BLASTX). Also, there are transcriptomic evidence supporting our findings. According to Bozdech and colleagues, this protein is expressed throughout the 48h intraerythrocytic cycle and is essential regarding virulence and host pathogen interactions (Bozdech et al., 2008). The partial sequences were also excluded from further analysis, since we cannot be certain of their origin (See Data & Methods).

**Table 14:** A summary table of the 9 unique proteins with no detectable homologs within *Plasmodium* or against NR database.

| A/A | PROTEIN ID | PLASMODB ID | LENGTH | ANNOTATION | CBRS | %CBRS | %GC | ND | BLASTP HITS | TBLASTN HITS | QIPP | STATUS | NOTES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | PFAL-IT-01-5413 | PFIT_bin07400 | 70 | erythrocyte membrane protein 1 (PfEMP1), pseudogene | Y | 11.43 | 27.23 | 0 | N/A | EF158073.1<br>AL844504.2<br>CP016995.1<br>AL844506.3<br>CP016997.1<br>AL844507.3<br>AL844503.2<br>CP016999.1<br>XM_002808758.1<br>AL844508.2<br>CP016994.1<br>AL844501.2<br>LN999943.1<br>CP016991.1<br>AL844505.2 | 0.33 | Excluded | Genome Not in NCBI database Wrong gene prediction |
| **2** | PVIV-Sal1-01-902 | PVX_089440 | 50 | hypothetical protein | K | 28.00 | 40.52 | 0 | N/A | N/A | 0.26 | Strain specific | PlasmoDB has transcriptomic data |
| **3** | PYOE-17XNL-01-1825 | PY01825 | 40 | hypothetical protein | N | 22.50 | 22.76 | 0 | N/A | N/A | 0.33 | Partial | Excluded due to partial sequence |
| **4** | PYOE-17XNL-01-3113 | PY03113 | 34 | hypothetical protein | K | 32.35 | 20.95 | 0.6 | N/A | N/A | 0.5 | Partial | Excluded due to partial sequence |
| **5** | PYOE-17XNL-01-3409 | PY03409 | 40 | hypothetical protein | K | 37.50 | 26.83 | 0.4 | N/A | N/A | 0.47 | Partial | Excluded due to partial sequence |
| **6** | PYOE-17XNL-01-4192 | PY04192 | 42 | hypothetical protein | K | 35.71 | 13.95 | 0.4 | N/A | N/A | 0.36 | Partial | Excluded due to partial sequence |
| **7** | PYOE-17XNL-01-4380 | PY04380 | 38 | hypothetical protein | N | 26.32 | 22.03 | 0.2 | N/A | XM_719579.1 | 0.42 | Partial | Excluded due to partial sequence |

| A/A | PROTEIN ID | PLASMODB ID | LENGTH | ANNOTATION | CBRS | %CBRS | %GC | ND | BLASTP HITS | TBLASTN HITS | QIPP | STATUS | NOTES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **8** | PYOE-17XNL-01-4963 | PY04963 | 37 | hypothetical protein | F | 29.73 | 18.42 | 0.1 | N/A | N/A | 0.23 | Partial | Excluded due to partial sequence |
| **9** | PYOE-17XNL-01-6944 | PY06944 | 34 | hypothetical protein | K | 23.53 | 29.52 | 0 | N/A | N/A | 0.18 | Partial | Excluded due to partial sequence |

The remaining 1192 proteins were divided into categories based on the number of BLASTP hits: 1002 produced only self-hits, thus qualifying for being labeled as unique/orphan/de novo and 190 proteins with more than 2 hits (Table 15). Interestingly, three *Plasmodium* species, namely *P. berghei* ANKA, *P. falciparum* 3D7 and *P. yoelii* 17X, do not have any unique protein. We believe that this observation is most probably due to the distribution of the available *Plasmodium* genomes across the genus phylogeny, but further work is necessary to rule out some biologically relevant explanations.

The Bayesian-inference *Plasmodium* species tree based on the *Plasmodium* core families indicates that the ancestor of the *P. yoelii* clade is *P. berghei* and that all three *P. yoelii are* closely related (Figure 18). The *P. falciparum* IT genome was derived from iterative mapping reads against *P. falciparum* 3D7 (Aurrecoechea et al., 2009) indicating that its unique proteins could be pseudogenes or gene prediction artefacts.

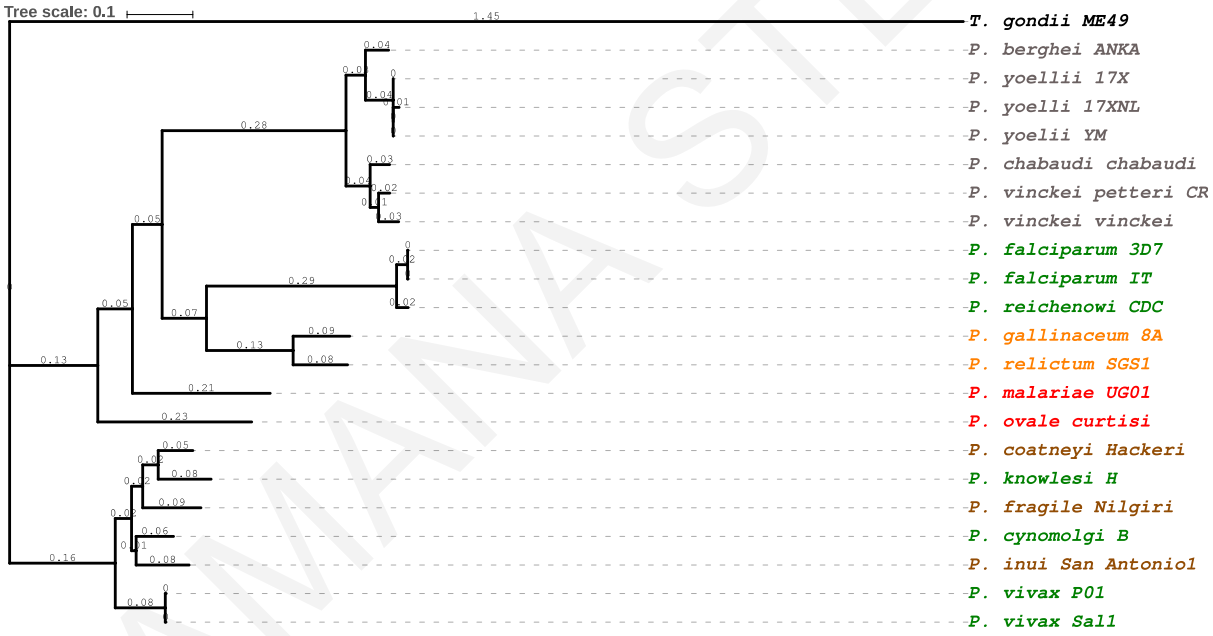



**Figure 18:** Bayesian-inference core genome phylogenetic tree of *Plasmodium* species. *Plasmodium* species are colored base on their currently known host: **Rodents**, **Simians/Humans**, **Simians**, **Humans** and **Avian**. Branches are labeled with their *length*. *Toxoplasma* gondii ME49 serves as the outgroup.

These observations are further supported by the Isolation Index of Organisms (IIO; (Fukuchi and Nishikawa, 2004)) value, where it seems to be higher when evolutionary close neighbors exist in the dataset (Table 15) as in *P. berghei* ANKA (IIO = -175.14) and *P. yoelii* 17X (IIO = -184.40) case. Since IIO is based on the logarithm of reciprocal best hits e-values (as defined in Data and Methods), a negative number with small absolute values suggests a

longer evolutionary distance between genomes. We observe that *Plasmodium* species of the rodent clade (which are genetically more similar according to the core pan-genome phylogeny) show lower IIO values. We could not identify any obvious correlation of the IIO measure and the number of unique proteins detected in each species.

**Table 15:** Summary of the results from the *all-vs-all* sequence comparisons against NR database for the unique proteins of *Plasmodium* pan-genome. The numbers in the *per species unique proteins* column are the pre-processed number of proteins.

| **Summary Table for the Plasmodia Unique proteins** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Total no. unique proteins** | 1201 | **No. excluded proteins:** | 1053 | **Total BLASTP hits:** | 1192 | **Total TBLASTN hits:** | 1089 | | |
| **Plasmodium Species** | **Index of Isolation Organisms (IIO)** | **Per species unique proteins** | | **No. BLASTP Hits** | | | | | |
| | | | | **1** | **2** | **3** | **4** | **5** | **6** |
| *P. berghei ANKA* | -175.14 | NO UNIQUE PROTEINS | | | | | | | |
| *P. chabaudi chabaudi* | -165.40 | 1 | | 1 | 0 | 0 | 0 | 0 | 0 |
| *P. coatneyi Hackeri* | -155.62 | 7 | | 7 | 0 | 0 | 0 | 0 | 0 |
| *P. cynomolgi B* | -137.12 | 32 | | 29 | 1 | 2 | 0 | 0 | 0 |
| *P. falciparum 3D7* | -150.09 | NO UNIQUE PROTEINS | | | | | | | |
| *P. falciparum IT* | -148.18 | 1 | | 0 | 0 | 0 | 0 | 0 | 1 |
| *P. fragile Nilgeri* | -169.82 | 86 | | 84 | 1 | 1 | 0 | 0 | 0 |
| *P. inui San Antonio1* | -170.11 | 380 | | 377 | 1 | 0 | 0 | 1 | 1 |
| *P. knowlesi H* | -158.07 | 4 | | 3 | 1 | 0 | 0 | 0 | 0 |
| *P. malariae* | -133.54 | 49 | | 15 | 19 | 14 | 1 | 0 | 0 |
| *P. ovale curtisi* | -117.28 | 69 | | 10 | 17 | 17 | 14 | 5 | 6 |
| *P. reichenowi CDC* | -147.99 | 6 | | 4 | 0 | 1 | 0 | 0 | 1 |
| *P. vinckei vinckei* | -164.73 | 7 | | 5 | 2 | 0 | 0 | 0 | 0 |
| *P. vinckei petteri* | -164.37 | 15 | | 9 | 3 | 1 | 2 | 0 | 0 |
| *P. vivax P01* | -126.65 | 9 | | 5 | 2 | 1 | 0 | 0 | 1 |
| *P. vivax Sal1* | -146.64 | 34 | | 12 | 14 | 2 | 4 | 2 | 0 |
| *P. yoelii 17XNL* | -192.30 | 492 | | 441 | 34 | 9 | 1 | 2 | 5 |
| *P. yoelii 17X* | -184.40 | NO UNIQUE PROTEINS | | | | | | | |
| *P. yoelii YM* | -190.23 | 1 | | 0 | 1 | 0 | 0 | 0 | 0 |

We also computed a non-homology based score, namely "Quality Index for Predicted Proteins" (QIPP score) to examine if this lack of homology is due to annotation artifact or due to sampling-bias in public genome databases (Wilson et al., 2007). According to Wilson G.A. and colleagues, there is increasing evidence that TRGs and orphan genes are of biological significance and hence, a growing need to disaffiliate authentic species/strain TRGs from erroneous gene predictions based on non-homology-based criteria (Wilson et al., 2007). They developed a method for bacterial and archaeal genomes of scoring CDS, namely QIPP scores, based on five criteria (Length, %CBRs, G+C content, Average amino acid cost and Neighborhood distribution) selected for their presumed ability to detect CDS which are unlikely to occur by chance (Wilson et al., 2007). Thus, we expect that longer CDS with typical nucleotide/amino acid composition (low A+T content), low %CBR (or no CBRs) are found in well conserved regions are more likely to encode proteins. QIPP scores with values close to zero would be the worst possible candidates for a real gene while, values close to one will indicate a protein coding gene. In Figure 19, we compare the QIPP score (computed either with or without the %CBR criterion) distribution between the pan-genome and unique proteins. We observe that, both in the pan-genome and unique proteins distribution the QIPP scores with %CBR begin at 0.35 while without %CBR at zero. Most QIPP scores for both distributions fall between 0.4 and 0.7 suggesting, a cut-off value for separating authentic functional proteins from gene prediction/annotation artifact to be set at 0.4.

However, the QIPP score was designed for bacterial and archaeal genomes which are mostly low to mild biased (both in terms of A+T content and %CBRs) genomes as opposed to the malaria parasites genomes. It's widely acknowledge that malaria parasites genomes are enriched in A+T and possess an unusually high incident of extremely biased proteins that play a critical role in host's cells invasion (Bozdech et al., 2008; Frech and Chen, 2013; Liu et al., 2006; Zilversmit et al., 2010) suggesting that *Plasmodium* QIPP score should be used with caution and in conjunction with other comparative genomic methods.

An extra step in our process of elimination of false positive hits was to carefully inspect each protein's self-hits (BLASTP and tBLASTN). All proteins annotated as 'partial cds' or 'partial mRNA' or 'fragment' were excluded from further analysis, leading to a final set of 148 out of the 1201 original set of putative unique protein sequences. When a protein sequence or mRNA is labeled as partial it usually means that the gene does not have all the start or stop

sequences at present or no UTR information (National Center of Biotechnology Information, 2017b) and thus, it suggest low quality sequence.

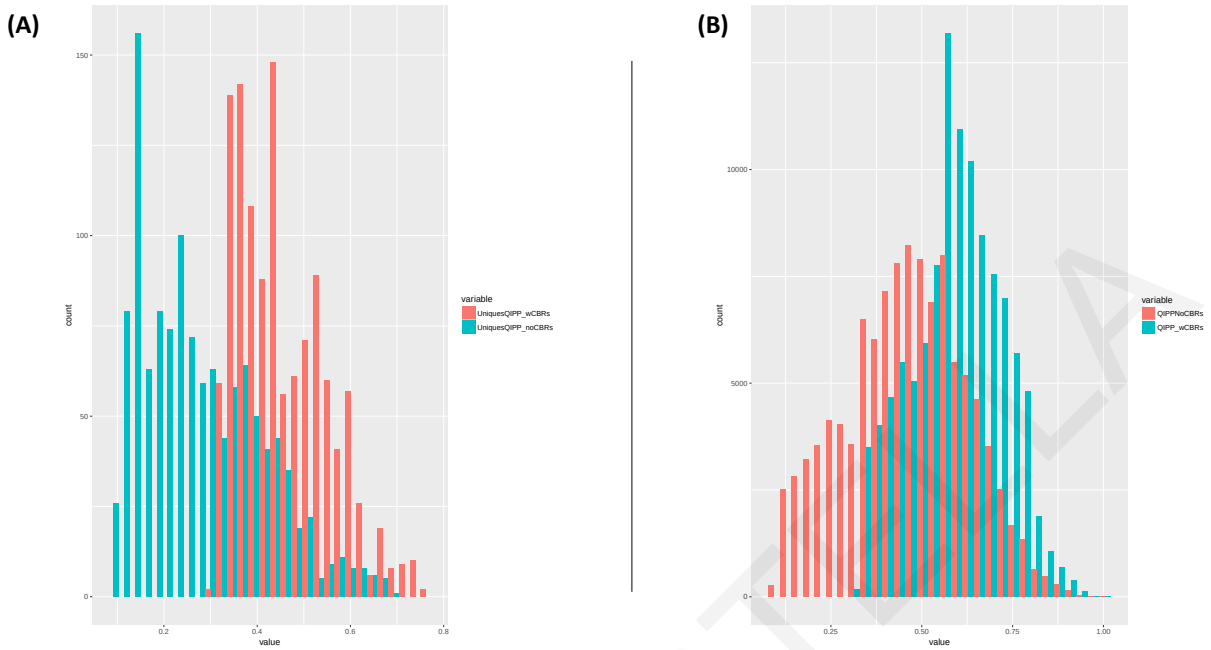



**Figure 19:** A histogram of the (A) unique and (B) pan-genome proteins QIPP scores distribution where salmon colored bars in both histograms denote the QIPP scores calculated *including* the *average %CBRs* and teal colored bars denote the QIPP scores calculated *without the average %CBRs* scores.

### Amino Acid Composition of Unique genes

The general belief is that orphan genes originated from spurious non-functional Open Reading Frames (ORFs) mainly because, these genes tend to be very short (~6 times shorter than mature genes), simpler in codon usage and amino acid composition (Arendsee et al., 2014) and that mostly encode intrinsically disordered proteins (Mukherjee et al., 2015; Verster et al., 2017; Wilson et al., 2017). Since malaria parasites genomes are characterized by high A+T content and thus, limited codon usage and amino acid composition, we wanted to test this hypothesis by comparing the average number of masked amino acids between the initially putative unique proteins versus the whole *Plasmodium* pan-genome proteins (as described in Data and Methods). We should note that for our pan-genome calculations we excluded the initially putative unique protein sequences as to avoid duplicated calculations.

The *Plasmodium* pan-genome is composed of 108,563 protein sequences where the average protein length is 665.43 and 62.8% of the proteins contain at least one CBR. In pan-genome, the most abundant residue types are Asparagine, Aspartic acid, Glutamic acid, Isoleucine, Leucine, Lysine and Serine (Figure 20) and while, these values serve as a descriptor of the

average global compositional bias of the proteins in malaria parasites proteome, it could be a source of CBR in protein sequences.

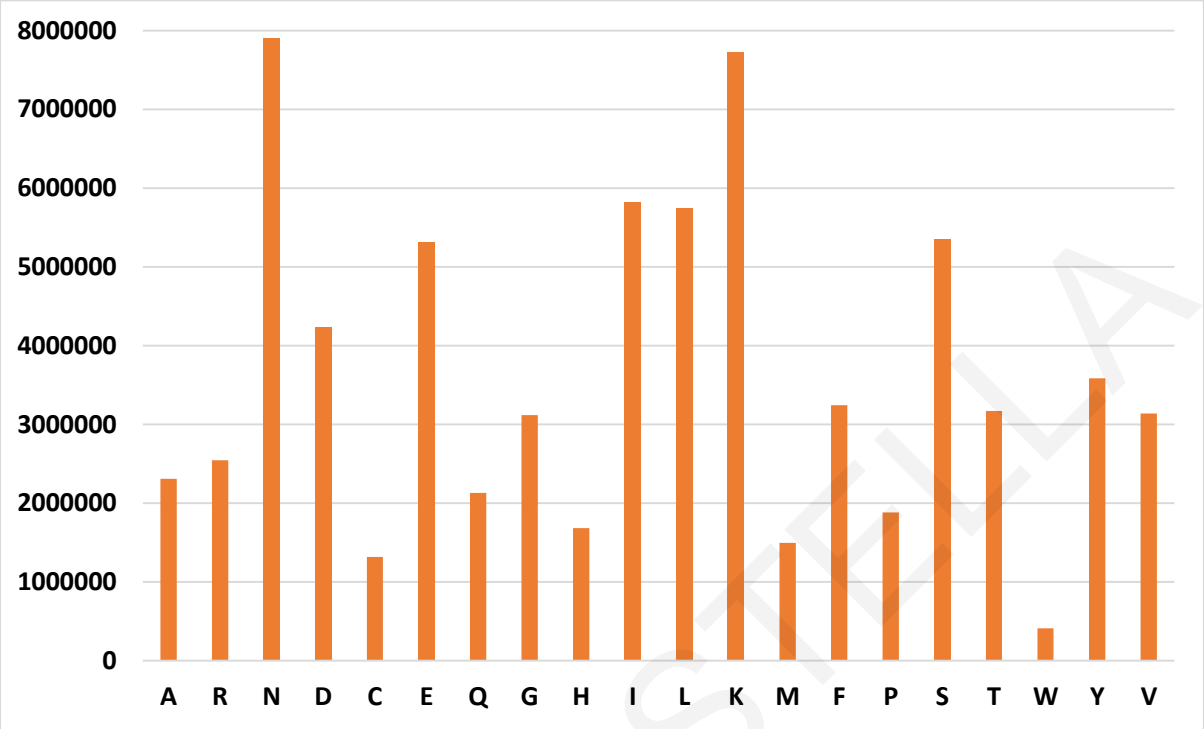



**Figure 20:** *Plasmodium* pan-genome amino acid background frequencies.

Although, one would expect that the most frequently masked residue types would follow the average global composition of *Plasmodium* pan-genome due to the excess of these residue types, yet, we observed that only Asparagine, Glutamic acid, Lysine and Serine being the amino acids contributing the most for *Plasmodium*'s heavily biased sequences (Figure 21).

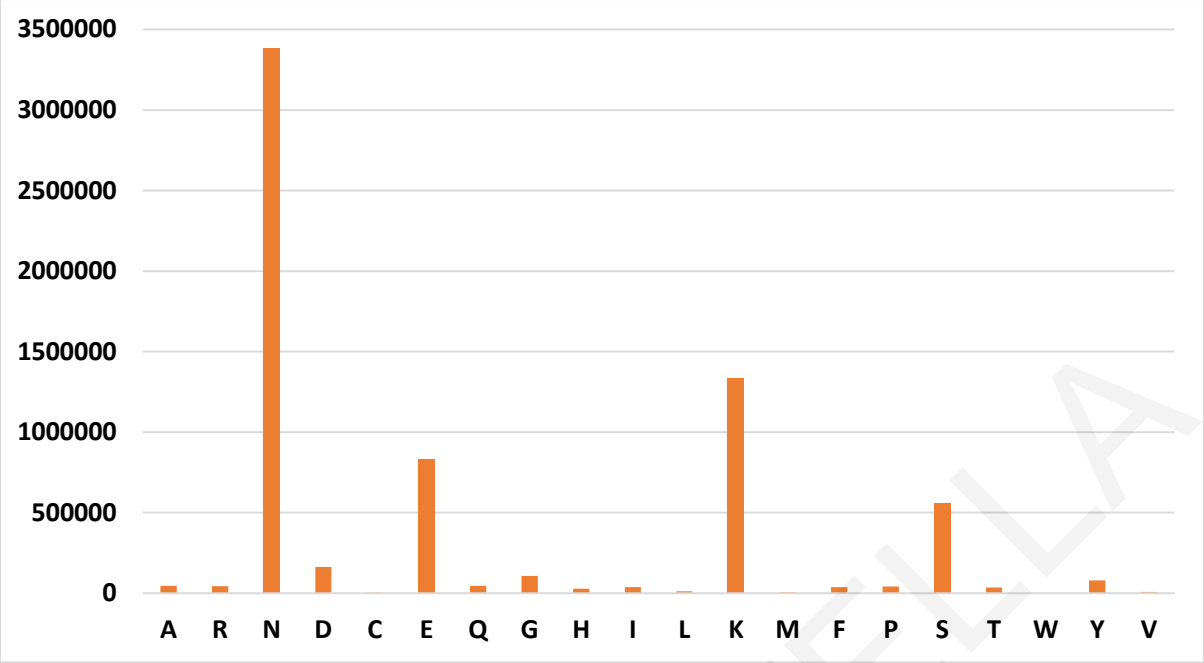



**Figure 21:** *Plasmodium* pan-genome masked residue types distribution.

In contrast, the 1201 putative unique proteins dataset has average protein length of 107.71 and 30.4% of the proteins contain at least one CBR. This, observation supports the hypothesis that unique genes are shorter and highly repetitive than non-unique as it appears the *Plasmodium* unique proteins are ~6 times shorter and almost a third of the proteins has compositional bias. We then, calculated for each of the twenty amino acids, the log fraction of the percentage of the pan-genome (excluding unique sequences; See Methods) masked amino acids against the percentage of the masked putative unique proteins' amino acids (Figure 22). This fraction will help us to determine which bias-causing residue types preferentially occur in putative unique protein sequences compared to the pan-genome and thus, help us determine possible evolutionary traits such as indications of de novo origination, or horizontal gene transfer (HGT).

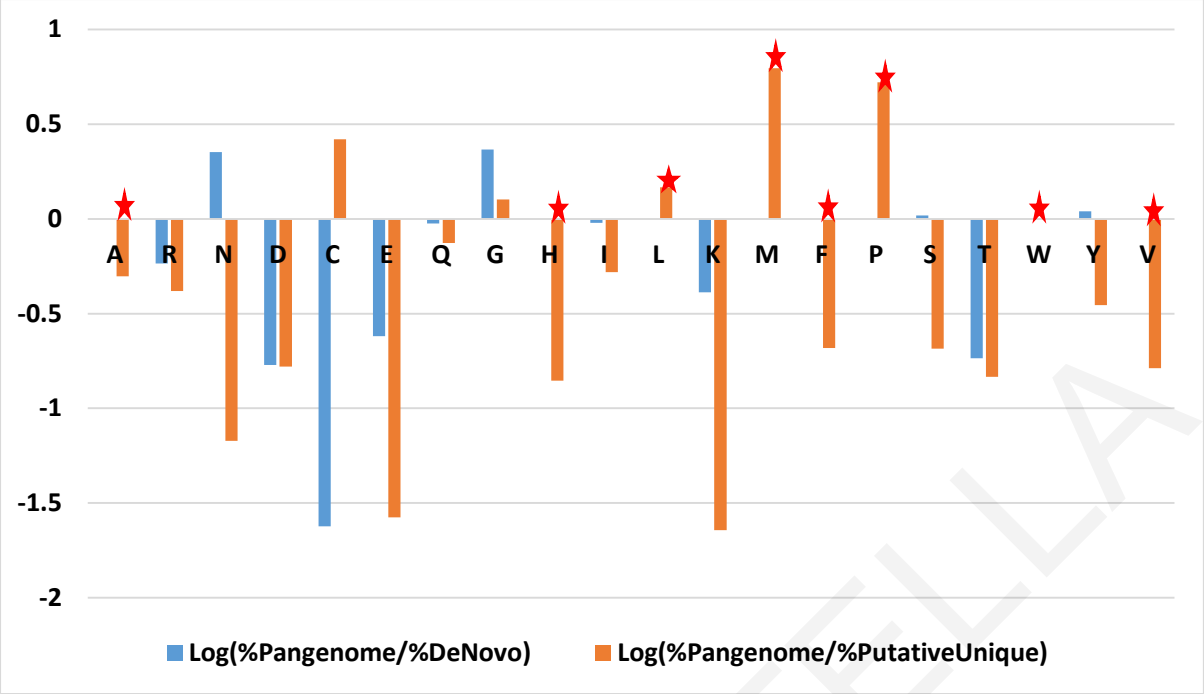



**Figure 22:** Log ratio bar chart of the percentage of the *Plasmodium* pan-genome (calculated by excluding the unique protein sequences from the calculations) *masked* versus the *unique and de novo's masked* residue types. The putative unique masked residue types were computed using the initial dataset (1201 proteins) as extracted from the MCL clustering file whereas, the *de novo's* masked residues were computed based on the final set of 96 proteins left after the post-processing analysis. With the star symbol, we mark residue types that the log fraction was undefined either in *de novo* or putative-unique masked amino acids distribution.

The most frequently masked residue types of putative unique proteins are Asparagine, Glutamic acid and Lysine (Figure 22). Remarkably certain residue types, such as Cysteine, Glycine, Methionine and Proline, display higher fraction of masked residues in pan-genome proteins in agreement with the notion that unique genes correspond to either recently evolved genes (i.e. young genes) or HGT that have not yet gone under the evolutionary forces to adapt to the species genome (McLysaght and Guerzoni, 2015; Wilson et al., 2005). Moreover, we computed the same log fraction for *de novo* proteins (i.e. verified by our post-processing analysis; see "Putative De Novo" section) where we observed that Arginine, Cysteine and Threonine to display higher fraction of causing bias residue types in *de novo* protein sequences rather than in pan-genome. This is particularly interesting, as all three residue types (Arginine, Cysteine and Threonine) are encoded by low A+T rich codons (with Adenine or Thymine holding the first and second position) which does not reflect the high A+T bias of *Plasmodium* genomes. Grantham and colleagues showed that each genome has a system for choosing between codons (Grantham et al., 1980) and the frequencies with which different synonymous codons (i.e. different codons coding the same amino acid) are

used for a particular amino acid, is called codon usage (Nørholm et al., 2012) . Codon usage varies greatly between different organisms but is shown that is specific both for the organism and the intracellular genome under investigation (Weiss et al., 2012). Frugier and colleagues suggested that Homopolymeric Repeat Regions (HPRs; an extreme case of CBRs) tend to behave like tRNA sponges in order to accumulate in co-translational folding and that provide a mechanism designed efficient enough to prevent co-translational miss-folding in *Plasmodial* proteins (Frugier et al., 2010). Thus, determining the differences between *Plasmodium* pan-genome and putative-unique codon usage we could assess the evolutionary forces that shape the unique genes and devise more efficient techniques for gene sequence optimization for structure and function determination.

For this reason, we computed a similar log ratio using *Plasmodium* pan-genome versus putative-unique codon usage (see Methods; Table 16). Interestingly, the Asn's codon AAT is more frequently observed in pan-genome coding sequences as opposed to the AAC codon that has higher usage by putative-unique CDS. Similar pattern seems to occur for Lys (AAA) codon which we observed higher codon usage in the pan-genome. In contrast, both Glu codons show higher frequency in putative-unique CDSs, however, the GAG codon is presenting higher usage in putative-unique CDSs compared to the A+T-richer GAA.

**Table 16:** Log ratio table of *Plasmodium* pan-genome versus putative-unique codon usage. We display the log fraction of codons computed as described in the respective Method section. Colored cells signify the most abundant residue types based on the global background frequencies of *Plasmodium* pan-genome as follows: **Asparagine**, **Aspartic acid**, **Glutamic acid**, **Isoleucine**, **Lysine**, **Serine** and the **stop** codons. N/A cells denote codons that were not observed in putative unique dataset. All values are rounded to 2 decimal points.

| First Letter | Second Letter: T | | | Second Letter: C | | | Second Letter: A | | | Second Letter: G | | | Third Letter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | TTT | Phe | 0.10 | TCT | Ser | N/A | TAT | Tyr | 0.10 | TGT | Cys | 0.04 | T |
| | TTC | | 0.01 | TCC | | -0.03 | TAC | | 0.12 | TGC | | 0.03 | C |
| | TTA | Leu | 0.14 | TCA | | -0.01 | TAA | Stop | -0.70 | TGA | Stop | -0.15 | A |
| | TTG | | 0.10 | TCG | | 0.02 | TAG | Stop | -0.75 | TGG | Trp | -0.72 | G |
| C | CTT | Leu | -0.02 | CCT | Pro | N/A | CAT | His | 0.02 | CGT | Arg | N/A | T |
| | CTC | | -0.05 | CCC | | -0.17 | CAC | | -0.25 | CGC | | -0.16 | C |
| | CTA | | -0.08 | CCA | | N/A | CAA | Gln | -0.07 | CGA | | N/A | A |
| | CTG | | 0.01 | CCG | | -0.19 | CAG | | -0.16 | CGG | | -0.26 | G |
| A | ATT | Ile | 0.09 | ACT | Thr | -0.03 | AAT | Asn | 0.23 | AGT | Ser | 0.09 | T |
| | ATC | | -0.12 | ACC | | -0.12 | AAC | | -0.02 | AGC | | 0.00 | C |
| | ATA | | 0.03 | ACA | | -0.13 | AAA | Lys | 0.05 | AGA | Arg | -0.16 | A |
| | ATG | Met | -0.10 | ACG | | -0.06 | AAG | | -0.08 | AGG | | -0.29 | G |

| | Second Letter | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | | | C | | | A | | | G | | | |
| G | GTT | Val | 0.02 | GCT | Ala | N/A | GAT | Asp | 0.18 | GGT | Gly | N/A | T |
| | GTC | | -0.15 | GCC | | -0.01 | GAC | | 0.06 | GGC | | -0.04 | C |
| | GTA | | -0.05 | GCA | | N/A | GAA | Glu | -0.02 | GGA | | N/A | A |
| | GTG | | -0.06 | GCG | | 0.00 | GAG | | -0.12 | GGG | | -0.06 | G |

Codons that are not frequently used as opposed to their synonymous are termed as *Rare Codons* (Nørholm et al., 2012) and earlier studies indicated that there is a correlation between codon usage and the expression levels (Bennetzen and Hall, 1982; Gouy and Gautier, 1982; Grantham et al., 1980). Research for mechanisms regarding rare codons and their effects on protein synthesis and cellular fitness, associated rare codons with two major driving forces of molecular evolution: mutation and natural selection (Duret, 2002; Hershberg and Petrov, 2008). The mutational explanation of rare codons arises from the properties of underlying mutational processes (Drummond and Wilke, 2009) while natural selection mechanisms state how rare codons influence the fitness of an organism (Plotkin and Kudla, 2011). Moreover, several studies indicate that synonymous codon changes can influence mRNA stability, mRNA structure, translational initiation, translational elongation and protein folding (Cannarozzi et al., 2010; Fredrick and Ibba, 2010; Kudla et al., 2009; Makino et al., 1997; Marin, 2008). Hence, the observation that the initially putative unique genes (1201) show a preference in rare codons provide an insight on the driving forces of unique candidate genes protein synthesis.

Another interesting observation is the higher stop (TAG and TAA) codon usage in putative-unique genes than in pan-genome (Table 16). This higher percentage of termination codons in putative unique coding sequences suggests that the majority of these genes are pseudogenes or at least, partial or low-quality coding sequences. Although, this observation seems to reinforce the hypothesis that unique genes are spurious sequences and should be eliminated as "junk" sequences, it could indicate that is the first step before codon adaptation. Pseudogenes and/or degenerated sequences often have premature termination codons (denoted by the star (*) character in protein sequences) that results in producing a non-functional protein . In general, pseudogenes may originate from a complete copy of the mRNA transcribed from the protein-coding gene, are only a partial copy of the corresponding mRNA or contain additional sequences than those expected to be present in

the mRNA (Vanin, 1985). Even though the DNA sequence of a pseudogene is very similar to its functional counterpart, variant mutations render the gene inactive (Vanin, 1985). Though, pseudogenization is considered as a neutral process and occasionally deleterious, it may shed light into a species evolutionary history by point into specific phenotypic changes and mutation rates. At the same time, it can help us understand the genetic basis of phenotypic evolution but regarding these findings, further research is needed to understand and evaluate their significance.

**"Putative unique proteins" with known annotations**

Among the initially unique candidates we noticed proteins annotated to belong to previously-defined protein families (e.g. 40S ribosomal protein S30) or as Major *Plasmodium* Families (e.g. PfEMP1). It's important to verify their annotation along with their unique status because most of these annotations resulted from automatic gene prediction pipelines and not based on experimental data.

Thus, we could provide evidence either supporting this annotation or suggest a more probable annotation. In this study, we observed 26 unique proteins with 'suspicious' annotations (Table 17) but most of these proteins were eliminated due to partial mRNA sequence in the post-processing step.

The first protein with dubious annotation is an 85-residue long *P. cynomolgi* protein (COGENT id: PCYN-B-01-1455) annotated as "*telomeric repeat binding factor 1*" (TRF1). Both BLASTP/TBLASTN (e-value: $10^{-6}$) sequence searches against *Plasmodium* pan-genome and NT/NR databases set this protein as unique. Consequently, we repeated the protein sequence comparisons using the integrated BLASTP/TBLASTN tools (with default parameters) in PlasmoDB website. The protein sequence comparison against PlasmoDBs Annotated Proteins confirmed its Orphan status while, the search against PlasmoDBs Annotated Genome found a statistically significant match (tBLASTN cut-off e-value: $2e^{-6}$) with a *P. fragile* hypothetical protein. Reviewing the alignment match with the PlasmoDBs Genome browser tool, we observed that the significant matches were partial matches to one exon of this protein (PlasmoDB id: AK88_02900). Then, using its genomic DNA sequence, we searched in PlasmoDB Annotated Transcripts (e-value threshold: $10^{-6}$) where we observed statistical significant matches to *P. chabaudi* (e-value: $9e^{-8}$), *P. falciparum* 3D7 (e-value: $7e^{-9}$), *P. falciparum* IT (e-value: $1e^{-11}$), *P. fragile* (e-value: $1e^{-17}$), *P. inui* (e-value: $3e^{-}$

[20]), *P. malariae* (e-value: $2e^{-9}$), *P. reichenowi* (e-value: $7e^{-9}$) and *P. vivax* (e-value: $4e^{-18}$). Based on the synteny map provided in PlasmoDB gene record, all these genes are syntenic orthologues annotated as TRF1 or hypothetical protein but there is a missing link between these genes and the *P. cynomolgi* gene. Furthermore, all genes are encoded by 5-6 exons while the *P. cynomolgi* putative TRF1 by 2 exons suggesting a deletion event or gene prediction artefact. Reviewing the individual pairwise alignments of each hit, we observed that only the *P. fragile*, *P. inui* and *P. vivax* hits matched the two *P. cynomolgi* exons and thus, we discarded the remaining hits as False Positives (FP).

Consequently, we retrieved the FASTA formatted protein sequences of *P. fragile*, *P. inui*, *P. vivax* and *P. cynomolgi* and set them as query against HMMER (Finn et al., 2015) batch on-line tool to search if the myb DNA binding domain is present (Ye et al., 2004) and thus, confirm the functional annotation. According to literature, the Telomeric Repeat-binding Factor (TRF) protein family is important for the regulation of telomere stability and bind directly to telomeric TTAGGG repeats via the myb DNA binding domain at the carboxyl terminus (Ye et al., 2004). Homologous sequences are apparent in most eukaryotic organisms but rapid divergence of the protein sequences leaves few traces of common ancestry (Horvath, 2013). A comparative genomics study in six *Plasmodium* species provided evidence that the *Plasmodium* TRF protein family is highly divergent as the *P. knowlesi* TRF1 protein sequence shares only 24% sequence similarity to the *P. falciparum* sequence (Frech and Chen, 2011).

Our HMMER search confirmed the myb DNA binding domain at the carboxyl terminus for all protein sequences except the *P. cynologi* protein, suggesting either a deletion of the myb DNA binding domain or gene prediction artefact. In **Figure** 23, we provide the MSA of these protein sequences (constructed on CLUSTAL Omega online version) where is evident that the *P. cynomolgi* protein TRF1 protein is gene prediction artifact.

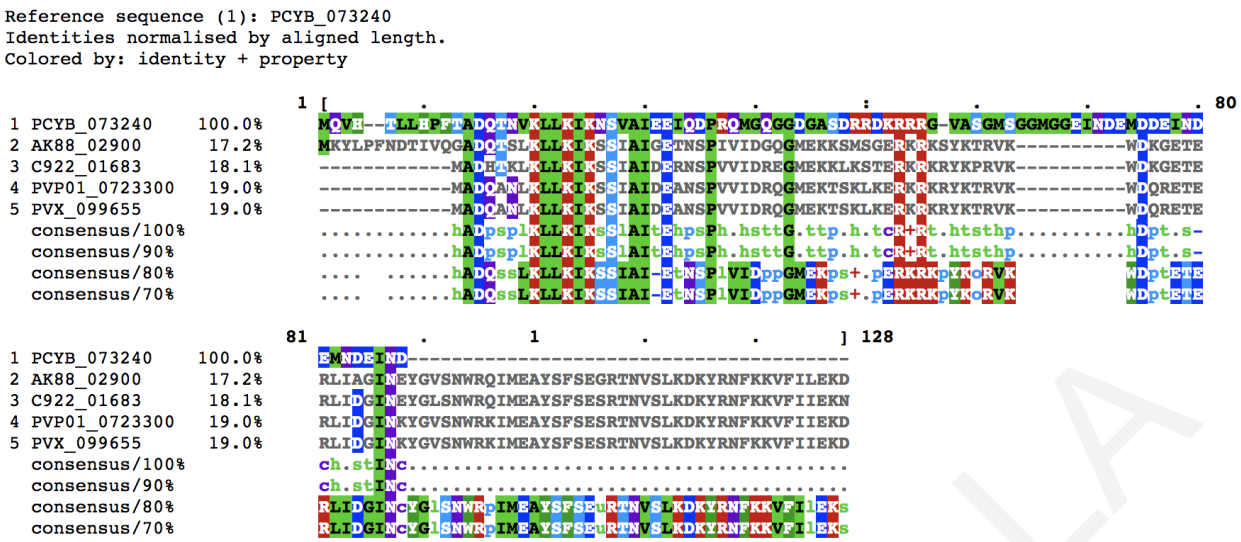



**Figure 23:** MSA of the *P. cynomolgi* TRF1 protein as constructed on the on-line version of Clustal Omega (Sievers et al., 2014) and visualized using mView (Li et al., 2015). PlasmoDB identifiers are displayed, corresponding to: PCYB_073240 - *P. cynomolgi*; AK88_02900 - *P. fragile* strain nilgiri; C922_01683 - *P. inui* San Antonio1; PVP01_0723300 - *P. vivax* P01; PVX_099655 - *P. vivax* Sal-1.

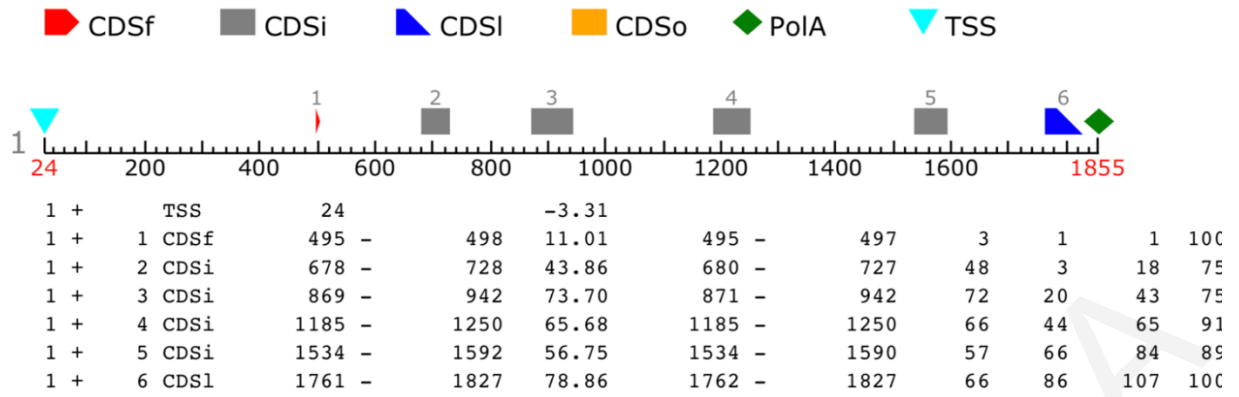
To provide further evidence that this sequence is a result of incorrect gene prediction, using the FGENESH+ (online gene prediction based on similarity tool; (Solovyev, 2007)), the TRF1 protein sequence of *P. vivax* Sal1 (PlasmoDB id: PVX_099655) as the protein templated for the gene prediction and the genomic sequence of PCYN-B-01-1455, we successfully predicted the complete mRNA and protein sequence of PCYN-B-01-1455. It's worth mentioning that the PCYN-B-01-1455 genomic sequence was retrieved from the genome browser tool of PlasmoDBs' where, we extracted a 2000bp nucleotide sequence (the predicted genomic sequence is 714bp) by selecting "Download Decorated FASTA File". In **Figure 24**, we observe that *P. cynomolgi* gene is actually 6 exons with the first exon starting at position 495 of the extracted genomic sequence and the last exon ends at 1827 confirming that the automatic gene prediction pipeline failed to correctly predict all exons. The sequence comparison (PlasmoDB BLASTP tool with default options) of the newly predicted protein sequence found statistically significant hits in all *Plasmodium* species currently deposited in PlasmoDB, annotated as TRF1.

**Table 17:** Summary table for the 26 proteins with suspicious annotation.

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | Accession No. | QIPP score | Status | Notes |
|---|---|---|---|---|---|---|---|---|---|
| **1** | PCYN-B-01-1455 | PCYB_073240 | 85 | telomeric repeat binding factor 1 | N/A | XP_004221769.1 | 0.31 | Artefacts | Wrong gene prediction |
| **2** | PCYN-B-01-1656 | PCYB_081840 | 36 | RAD protein | N/A | XP_004221970.1 | 0.28 | Artefact | Functional prediction artefact |
| **3** | PCYN-B-01-2071 | PCYB_092240 | 78 | cyclophilin | N/A | XP_004222385.1 | 0.1 | Partial | Excluded due to partial sequence |
| **4** | PCYN-B-01-2432 | PCYB_101490 | 105 | SICAvar-like protein | N/A | XP_004222746.1 | 0.54 | Partial | Excluded due to partial sequence |
| **5** | PCYN-B-01-3808 | PCYB_127400 | 213 | KS1 protein precursor | D, E, K | XP_004224122.1 | 0.44 | TRG | Missed due to CBRs and strict e-value |
| **6** | PCYN-B-01-4079 | PCYB_133200 | 153 | ABC transporter | N/A | XP_004224393.1 | 0.34 | Artefact | No Pfam domain Wrong gene prediction/functional annotation |
| **7** | PCYN-B-01-5439 | PCYB_005550 | 102 | CYIR protein | N/A | XP_004228024.1 | 0.12 | Partial/Fragment? | Probably partial? |
| **8** | PCYN-B-01-5000 | PCYB_001150 | 96 | CYIR protein | N/A | XP_004228313.1 | 0.26 | Partial | Excluded due to partial mRNA sequence |
| **9** | PCYN-B-01-5246 | PCYB_003620 | 57 | CYIR protein | N/A | XP_004227831.1 | 0.19 | Partial | Excluded due to partial mRNA sequence |
| **10** | PCYN-B-01-545 | PCYB_041140 | 32 | 40S ribosomal protein S30 | N/A | XP_004225313.1 | 0.18 | Artefact | Wrong functional prediction |
| **11** | PCYN-B-01-5513 | PCYB_006290 | 117 | CYIR protein | N/A | XP_004228098.1 | 0.13 | Partial | Excluded due to partial sequence |
| **12** | PFAL-IT-01-4597 | PFIT_1401900 | 168 | erythrocyte membrane protein 1, PfEMP1, putative | N | Not in NR database | 0.37 | Pseudogene | PlasmoDB confirms BLASTP/tBLASTN hits No Pfam domain |

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | Accession No. | QIPP score | Status | Notes |
|---|---|---|---|---|---|---|---|---|---|
| **13** | PMAL-UG01-01-1041 | PmUG01_05027600 | 716 | Plasmodium exported protein (PHIST), unknown function | E, K | SBS96832.1 SBT87413.1 SBT74945.1 | 0.46 | FN | Pfam confirms functional annotation Missed due to CBRs and strict e-value |
| **14** | PMAL-UG01-01-1166 | PmUG01_05040000 | 209 | asparagine-rich protein | N, D | SBS94918.1 SBT75071.1 | 0.29 | FN | Missed due to CBRs |
| **15** | PMAL-UG01-01-3168 | PmUG01_10046900 | 364 | liver stage antigen 1, putative | E, N, K | SBT01459.1 | 0.3 | Artefact | No Pfam match No significant similarity to reviewed PF3D7 (A0A143ZZD7) LSA Wrong functional annotation PlasmoDB OG5_141754 |
| **16** | PMAL-UG01-01-6423 | PmUG01_00075300 | 31 | fam-l protein | N/A | SBT85994.1 | 0.18 | Partial | Excluded due to partial sequence |
| **17** | PMAL-UG01-01-5279 | PmUG01_14029200 | 61 | apical ring associated protein 1, putative | N/A | SCP03238.1 SBT80652.1 | 0.33 | Wrong gene prediction Correct coordinates: complement (919943-920123) | No Pfam domain PlasmoDB tBLASTN hits |
| **18** | PMAL-UG01-01-6543 | PmUG01_API001300 | 32 | ribosomal protein L23, putative | N/A | YP_009307909.1 | 0.22 | Partial | Excluded due to partial sequence |
| **19** | POVA-CURT-01-5887 | PocGH01_00160000 | 418 | PIR protein | S, I, E | SBT30951.1 | 0.49 | TRG | PlasmoDB (BLASTP/tBLASTN) Missed due to CBRs PlasmoDB orthologous |

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | Accession No. | QIPP score | Status | Notes |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | group: OG5_244515 |
| **20** | POVA-CURT-01-2511 | PocGH01_10038700 | 497 | merozoite surface protein 3, putative | E, A, N | SBT73672.1 | 0.26 | Fragment | Match to P. falciparum protein (PF10_0345) PlasmoDB: OG5_126560 |
| **21** | PREI-CDC-01-4428 | PRCDC_1369500 | 64 | erythrocyte binding like protein 1, fragment | N/A | XP_012765062.1 | 0.16 | Fragment | Excluded due to protein fragment |
| **22** | PREI-CDC-01-5648 | PRCDC_0054800 | 99 | erythrocyte membrane protein 1, PfEMP1, putative | N/A | XP_012760598.1 | 0.1 | Partial | Excluded due to mRNA partial sequence |
| **23** | PVIV-P01-01-5702 | PVP01_0004380 | 488 | PIR protein | S | SCO70940.1 | 0.26 | Partial | Excluded due to mRNA partial sequence |
| **24** | PVIV-P01-01-5773 | PVP01_0010540 | 76 | gamma-glutamylcysteine synthetase, putative | N/A | SCA82328.1 | 0.2 | Partial | Excluded due to mRNA partial sequence |
| **25** | PVIV-P01-01-5985 | PVP01_0001340 | 80 | PIR protein, fragment | N/A | SCA82767.1 | 0.17 | Partial | Excluded due to mRNA partial sequence |
| **26** | PVIV-Sal1-01-3894 | PVX_118010 | 84 | Histone H1, gonadal, putative | K | XP_001615894.1 | 0.32 | Partial | Excluded due to mRNA partial sequence |




| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | + | | TSS | 24 | | -3.31 | | | | | |
| 1 | + | 1 | CDSf | 495 | 498 | 11.01 | 495 | 497 | 3 | 1 | 1 | 100 |
| 1 | + | 2 | CDSi | 678 | 728 | 43.86 | 680 | 727 | 48 | 3 | 18 | 75 |
| 1 | + | 3 | CDSi | 869 | 942 | 73.70 | 871 | 942 | 72 | 20 | 43 | 75 |
| 1 | + | 4 | CDSi | 1185 | 1250 | 65.68 | 1185 | 1250 | 66 | 44 | 65 | 91 |
| 1 | + | 5 | CDSi | 1534 | 1592 | 56.75 | 1534 | 1590 | 57 | 66 | 84 | 89 |
| 1 | + | 6 | CDSl | 1761 | 1827 | 78.86 | 1762 | 1827 | 66 | 86 | 107 | 100 |

**Figure 24:** The PCYN-B-01-1455 gene prediction by sequence similarity where, the number of predicted exons is six starting at 498bp and ends at 1827bp at positive strand (genomic location: DF157099: 996577..998576 (+)) of the submitted genomic sequence. The *P. vivax* Sal1 (PVX_099655) protein sequence was used as the required protein sequence template.

Furthermore, we confirmed, through Pfam protein sequence analysis online tool, that this predicted sequence belongs in the TRF1 protein family (Figure 25). Thus, we provide the correct genomic locus, predicted mRNA and protein sequence of the *P. cynomolgi* B TRF1 protein.

Pfam Matches — Advanced

| | Family | | Clan | Description | Cross-refs | Start | End | Domain E-values | |
|---|---|---|---|---|---|---|---|---|---|
| | Id | Accession | | | | | | Ind. | Cond. |
| > | Myb_DNA-binding | PF00249.30 | CL0123 | Myb-like DNA-binding domain | | 49 | 98 | 3.7e-07 | 4.5e-11 |

**Figure 25:** Pfam sequence analysis of the predicted by similarity *P. cynomolgi* TRF1 protein sequence.

The RAD protein (also known as "Pv-fam-e" protein) was first detected in *P. vivax* genome as part of an eight-gene family (Pv-fam-a to -e and -g to -i) (Carlton et al., 2008). Known orthologues of the RAD protein in Pfam database are found in all *Plasmodium* primate-infecting clade, however, the *P. cynomolgi* RAD protein sequence (COGENT id: PCYN-B-01-1656) was clustered as Unique. In order to address this issue, we first performed BLASTP/tBLASTN sequence searches in PlasmoDB Annotated Proteins/Genome respectively (with default options). Both searches confirm the unique status as the only statistically significant match was itself). Then, using the genomic DNA sequence of PCYN-B-01-1656 we performed gene prediction by similarity using the FGENESH+ (Solovyev, 2007) online tool and using as a protein template the *P. vivax* Sal-1 RAD protein (UniProt/TrEMBL id: A5K5Y3). However, no reliable

predictions were feasible suggesting that this sequence is probably functional annotation artefact.

Another case of a unique protein that based on its annotation it belongs to this protein family is a *P. malariae* protein (COGENT id: PMAL-UG01-01-3168). This protein's annotation is "liver stage antigen 1, putative" (also known as Pv-fam-g) located in a poorly syntenic neighborhood (ND = 0.1) and its amino acid sequence is heavily biased (35.71 %CBRs). Both protein sequence searches in *Plasmodium* pan-genome and NR database set it as unique but when we searched in PlasmoDB (using both the umasked and masked protein sequence) we observed weak matches with *P. ovale* (e-value: 0.014), *P. vivax* (e-value: 0.013), *P. inui* (e-value: 0.033) and *P. knowlesi* (e-value: 0.55). Inspection of their pairwise alignment showed local similarity either in the N-terminal or C-terminal sequence. Then, we retrieve the *P. falciparum* 3D7 LSA1 protein sequence (UniProtKB id: A0A143ZZD7) from UniProtKB/Swiss-Prot and performed pairwise alignment using the BLASTP online tool. The pairwise alignment showed a similarity pattern at the N-terminal (max query cover: 48%, e-value: 0.0023 and 27 %identity) between the *P. malariae* and *P. falciparum* sequence (UniProtKB id: A0A143ZZD7, protein length: 1162) but this observation does not provide enough evidence that the two sequences share a common ancestor or are homologues. In addition, the protein sequence search in Pfam did not show any indication of the existence of known protein domains. Thus, we believe the annotation is a function prediction artefact.

The *SICAvar* is one of the major Variant Surface Antigen (VSA) gene families (28 genes) found in *P. knowlesi* genome that demonstrated to undergo antigenic variation (Al-Khedery et al., 1999; Frech and Chen, 2013; Pain et al., 2008). To date SICAvar genes were, also, described, in *P. fragile* (Pfam number of sequences: 235) and in *P. cynomolgi* (Pfam number of sequences: 1). Thus, the *P. cynomolgi* protein (COGENT id: PCYN-B-01-2432) annotated as "SICAvar-like protein" could be a pseudogene, another gene of this family or a functional prediction error. Even though, we performed a Pfam protein sequence search where we confirm the functional domain of "SICA C-terminal inner membrane domain" (Pfam e-value: $2.5e^{-09}$) we set this protein as false positive due to partial mRNA sequence.

Following similar methodology for the "KS1 protein precursor" *P. cynomolgi* (COGENT id: PCYN-B-01-3808) protein we observed protein sequence similarity (masked sequenced BLASTP in PlasmoDB Annotated Proteins with default settings) to KS1 protein precursor from *P. knowlesi* ($8e^{-04}$) and *P. vivax* ($5e^{-04}$). PlasmoDB gene records for these protein precursors cluster them in Orthologue group OG5_133023 along with uncharacterized/hypothetical proteins from almost all *Plasmodium* species (with the exception of *P. malariae* and *P. vinckei*) that share structural similarities to Protein Data Bank entries (PDB; (Berman et al., 2000)) annotated as "Cytochrome b-c1 complex subunit 6" and "mTERF domain-containing protein 2". Furthermore, our Neighborhood Distribution score (ND = 0.9) show that are found in a highly conserved syntenic neighborhood, indicating they are essential genes. A possible explanation why this protein was clustered as unique could be the combination of heavy masking (72.77 %CBR) and the strict e-value cutoff we set.

Members of ABC transporter protein family belong to the ATP-Binding Cassette superfamily, which are water-soluble transmembrane domains that utilize ATP to translocate a variety of compounds across biological membranes (Hung et al., 1998). It can be found in all living organisms and its amino acid sequence (approximately 200 residues long) is highly conserved but with unrelated architectures of the transmembrane domains (Hollenstein et al., 2007). In *Plasmodium* species, numerous studies demonstrated the significance of ABC transporter proteins in drug-resistance mechanisms (Kavishe et al., 2009; Peel, 2001; Veiga et al., 2011). To date, three ABC transporter genes and numerous polymorphisms were characterized in *P. falciparum* genome (Okombo et al., 2013; Veiga et al., 2011). To confirm that *P. cynomolgi* ABC transporter protein belongs to this superfamily and it's not a faulty gene/functional prediction, we first perform BLASTP search in PlasmoDB Annotated proteins (with default settings) where we confirm its unique status. The tBLASTN search (also with default settings) against PlasmoDB Annotated genomes, however, returned statistically significant matches in all *Plasmodium* species. Inspecting the tBLASTN matches pairwise alignments and synteny map, we observed that the *P. cynomolgi* gene is in a highly conserved syntenic neighborhood (ND = 0.9) but it's only a partial match to the *P. falciparum* ABC-transporter gene. The tBLASTN matches from other Plasmodium species were also partial to uncharacterized/hypothetical proteins. The Pfam

search for this protein did not found any statistically significant match indicating that probably is annotation artifact of the automatic pipeline.

The CYIR protein family belongs to one of the largest variant gene families predicted to be involved in antigenic variation conserved throughout the *Plasmodium* phyla. It was first described in *P. vivax* (the vir family (del Portillo et al., 2001)) and later in *P. yoelii* (the yir family (Carlton et al., 2002)), *P. berghei* (the bir family (Janssen et al., 2002)), *P. chabaudi* (the cir family; (Janssen et al., 2002)), *P. knowlesi* (the kir family; (Janssen et al., 2004)) and *P. cynomolgi* (the cyir family (Tachibana et al., 2012)). Distant relatives of these protein families, also known to be involved in antigenic variation, is the *P. falciparum* PIR family sharing 20-30% sequence similarity (del Portillo et al., 2001). Jansen C.S. and colleagues carried out a cross-species phylogenetic analysis using 157 amino acid PIR sequences from *P. vivax*, *P. knowlesi*, *P. yoelii*, *P. berghei* and *P. chabaudi,* suggesting that PIR genes share an ancestral gene sequence, although its unknown whether any functional, regulatory or export mechanisms have been conserved (Janssen et al., 2004). Originally, it was believed to be part of the SICAvar family (Janssen et al., 2002) but conclusive evidence that these VSA families are evolutionarily linked through a conserved intracellular domain were obtained almost a decade after their characterization (Frech and Chen, 2013). Through our analysis of the *Plasmodium* pan-genome putative unique proteins, we observed, four *P. cynomolgi* entries annotated as "CYIR protein" and one *P. ovale* annotated as "PIR protein". For three CYIR proteins, we confirmed the functional annotation through Pfam protein sequence search (**Table** 17). These three proteins have partial mRNA sequences and were excluded from further analysis while for the one (COGENT id: PCYN-B-01-5439) with complete predicted mRNA sequence we performed additional protein and mRNA sequence searches in PlasmoDB Annotated Proteins/Genomes (BLASTP/tBLASTN/BLASTN with default options) as to confirm its annotation. Statistically significant sequence similarity was observed with multiple hits annotated as CYIR/VIR between *P. cynomolgi* and *P. vivax* proteins and genomes respectively. These results indicate that PCYN-B-01-5439 belongs to the CYIR protein family but it is probably a partial sequence and that is why it is clustered as unique. The *P. ovale* PIR was labeled as TRG because significant sequence similarity was observed only with *P. cynomolgi* (e-value: $3e^{-4}$) and *P. vivax* (e-value: $4e^{-6}$).

Another intriguing case is one of the *P. cynomolgi* proteins, which is a 78 amino acids peptide annotated as cyclophilin. In literature, cyclophilins are described as a family of proteins that bind to the immunosuppressive drug ciclosporin A (CsA) and, *in vitro*, assist in protein folding (Stamnes et al., 1992). They are widely distributed among different organisms, including *Plasmodium* (Marín-Menéndez et al., 2012) . Significant sequence similarity was observed in genome level between this protein and all other *Plasmodium* species in our dataset and in species from other genera. The annotation from most tBLASTN hits (including its own mRNA) for this protein is cyclophilin partial mRNA or peptidyl-prolyl cis-trans isomerase-like providing further evidence for its annotation. However, when we searched this protein in Pfam (Finn et al., 2016) no statistically significant match was found suggesting that either its missing the peptidyl-prolyl cis-trans isomerase domain due to partial mRNA sequence or it's a product of incorrect gene prediction.

The *P. cynomolgi* protein (COGENT id: PCYN-B-01-545) annotated as "40S ribosomal protein S30" is another case of incorrect gene prediction. Using the protein sequence of the *P. falciparum* "40S ribosomal protein S30" from UniProtKB/SwissProt (UniProt id: O96269) we performed a tBLASTN (PlasmoDB tBLASTN tool with default options) search against the *P. cynomolgi* genome. The pairwise alignment revealed that this unique protein is a case of incorrect gene prediction (Figure 26) as the predicted gene location is at 58237 - 59599 (-) of the DF157096 chromosome while, the significant match is in 58768-58598 (-) of the same chromosome.

```
 Link to Genome Browser,   Strand = > DF157096 | organism=Plasmodium_cynomolgi_strain_B | version=2012-09-19
| length=801051 | SO=chromosome
Length=801051

 Score =  100 bits (248),  Expect = 1e-25, Method: Compositional matrix adjust.
 Identities = 50/57 (88%), Positives = 54/57 (95%), Gaps = 0/57 (0%)
 Frame = -3

Query  1      MGKVHGSLARAGKVKNQTPKVPKLDKKKRLTGRAKKRQLYNRRFSDNGGRKKGPNSK  57
              +GKVHGSLARAGKVKNQTPKVPK+ K+K+LTGRAKKRQLYNRRFSD  GRKKGPNSK
Sbjct  58768  VGKVHGSLARAGKVKNQTPKVPKVSKRKKLTGRAKKRQLYNRRFSDGTGRKKGPNSK  58598
```

**Figure 26:** The pairwise alignment between the reviewed *P. falciparum* 40S ribosomal 30S protein sequence and the unique *P. cynomolgi* protein.

Additionally, we performed gene prediction by similarity (FGENESH+; (Solovyev, 2007)) using the PCYN-B-01-545 genomic sequence (from PlasmoDB's gene record) and *P. falciparum* 3D7 reviewed protein sequence (UniProt id: O96269) as the required protein template (Figure 27).

This predicted gene is composed of 1 exon, matching the coordinates of the tBLASTN statistically significant hit we observed above.

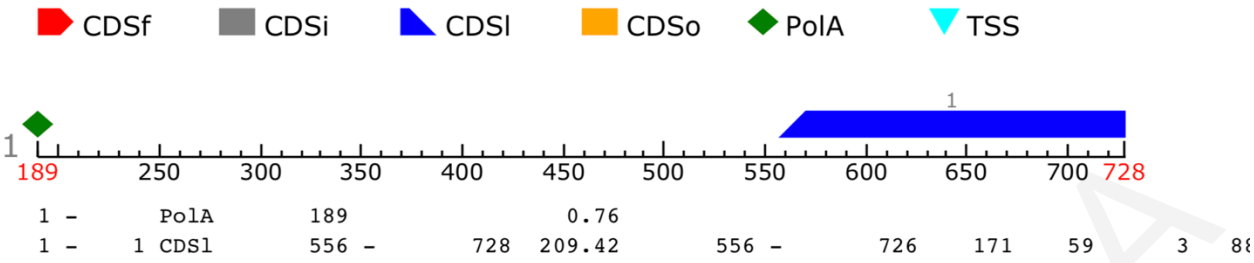



**Figure 27:** The *P. cynomolgi* "40S ribosomal protein S30" gene prediction by similarity where it predicts one exon starting at 556bp and ends at 728bd at the minus strand of the submitted genomic sequence (genomic location: DF157096: 58037..59799 (-)). The gene prediction was performed by the FGENESH+ online tool (Solovyev, 2007).

Additionally, the predicted protein sequence is 56 amino acids long as opposed to the 32 residues long PCYN-B-01-545 while, Pfam protein sequence analysis (e-value: $6.1e^{-25}$) confirms the functional annotation of "40S ribosomal protein S30" (Figure 28).

Pfam Matches — Advanced

| | Family | | Clan | Description | Cross-refs | Start | End | Domain E-values | |
|---|---|---|---|---|---|---|---|---|---|
| | Id | Accession | | | | | | Ind. | Cond. |
| > | Ribosomal_S30 | PF04758.13 | n/a | Ribosomal protein S30 | | 1 | 55 | 6.1e-25 | 3.7e-29 |

Your search took: 0.01 secs

**Figure 28:** Pfam protein sequence analysis of the predicted protein of PCYN-B-01-545.

For the *P. falciparum* IT protein annotated as "Erythrocyte membrane protein 1, PfEMP1, putative", which is the best-studied and clinically most relevant example of parasite-host interactions (Craig and Scherf, 2001; Frech and Chen, 2011), our analysis was performed in Pfam database and PlasmoDB BLASTP/tBLASTN tools since its genome is not currently deposited in NT/NR databases (Genome and National Center for Biotechnology Information, 2018). The PfEMP1 protein family is encoded by approximately 60 members of the *P. falciparum*-specific var gene family and its primary structure is composed of three structural domains: an extracellular domain (ECD), a transmembrane domain (TMD) and an intracellular acidic terminal segment (ATS) (Bull and Abdi, 2016; Frech and Chen, 2011). The top hits in *P. falciparum* 3D7 annotated proteins are "PfEMP1 pseudogene" and our protein sequence search in Pfam database did not produce any significant match with either PfEMP1 domains currently

deposited in Pfam. Furthermore, its gene is in a poorly conserved syntenic neighborhood (ND = 0) indicating that it's probably a pseudogene and excluded from further analysis.

As an Orphan was classified the *P. malariae* "Plasmodium exported protein (PHIST)" due to heavy masking (20.81 %CBRs) and strict e-value threshold. The PHIST gene family encodes more than 80 proteins of unknown function and localization within the host cell (Sargeant et al., 2006) that share a six-helical *Plasmodium* RESA N-terminal (PRESAN) domain (Sargeant et al., 2006; Warncke et al., 2016). The Pfam protein sequence search using the masked protein as query found statistically significant match to the PRESAN domain (Pfam e-value: $2.5e^{-8}$) confirming its functional annotation.

Another example of false positive unique protein due to CBRs, is the *P. malariae* "Asparagine-rich protein" (24.4 %CBRs) which, in fact, has single-copy orthologues in almost all malaria parasites (except in *P. gallinaceum* and *P. vinckei* vinckei) currently deposited in PlasmoDB. The Asparagine (N)-rich CBRs are one of the most abundant CBR types of *Plasmodium* species found in almost all *Plasmodium* protein families (Muralidharan and Goldberg, 2013; G. P. Singh et al., 2004). Several studies suggested that these repeats could act as tRNA sponges (Frugier et al., 2010), assist in protein-protein interaction, immune evasion and antigenic variation (Hughes, 2004; Karlin et al., 2002; Verra and Hughes, 1999).

Similar analysis was followed for the *P. malariae* "apical ring associated protein 1" (COGENT id: PMAL-UG01-01-5279). This protein is in a conserved syntenic neighborhood (ND = 0.9) in *P. malariae* apicoplast genome indicating that probably is one of the essential post-transcriptional processing genes for *Plasmodium* species (Arisue et al., 2012; Nisbet and McKenzie, 2016; Sato et al., 2013; Zuegge et al., 2001). The search for genes encoding similar proteins as the *P. malariae* one, returned statistically significant match to a *P. gallinaceum* transcript (e-value: $1e^{-4}$), which encodes a protein belonging in the same orthologue group (OG5_pber|PBANKA_1410950) with the experimentally characterized *P. berghei* AP2 protein (Kaneko et al., 2015). We retrieved the protein sequences of the orthologous group OG5_pber|PBANKA_1410950 in FASTA format and constructed an MSA using the online version of CLUSTALW (Sievers et al., 2014). The MSA (Figure 29) indicated that *P. malariae* protein sequence shares very little sequence similarity with the other protein sequences.

Using its genomic DNA sequence, we performed BLASTN search against Annotated transcripts where we observed statistically significant matches only with *P. relictum* (e-value: $2e^{-26}$), *P. ovale* (e-value: $1e^{-18}$) and *P. gallinaceum* (e-value: $6e^{-15}$) genomes. Then, we performed a BLASTX search (translated nucleotide to protein sequence), confirming the BLASTN hits, but in a different starting reading frame (Figure 30). This finding suggests gene prediction artefact.

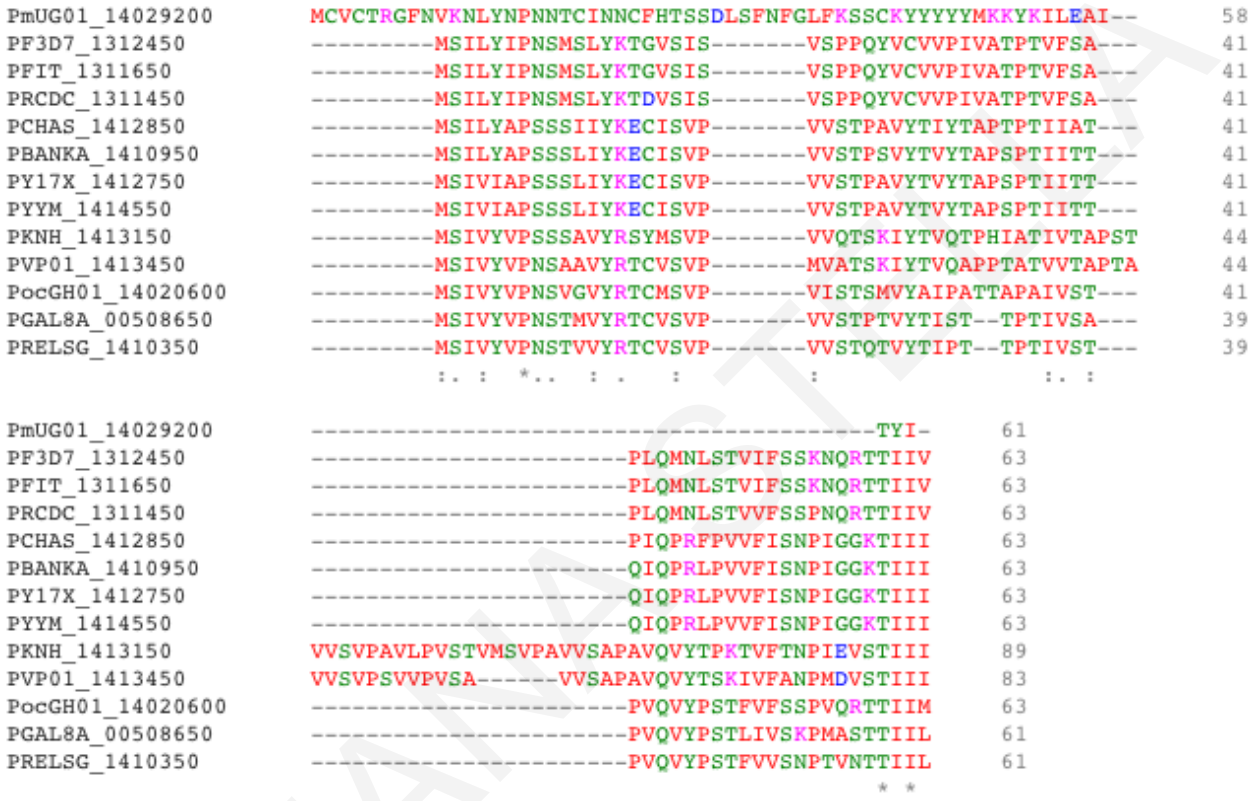



**Figure 29:** MSA of the "Apical ring associated protein 1" orthologue group OG5_pber|PBANKA_1410950.

```
> PRELSG_1410350.1-p1 | transcript=PRELSG_1410350.1 | gene=PRELSG_1410350
| organism=Plasmodium_relictum_SGS1-like | gene_product=apical
ring associated protein 1, putative | transcript_product=apical
ring associated protein 1, putative | location=PRELSG_14_v1:380333-380518(-)
| protein_length=61 | sequence_SO=chromosome
| SO=protein_coding
Length=61

 Score = 59.7 bits (143),  Expect = 2e-12, Method: Compositional matrix adjust.
 Identities = 35/48 (73%), Positives = 40/48 (83%), Gaps = 2/48 (4%)
 Frame = +2

Query  2    CVSVPAVSTSRIYTIPTTPVSTIVSTPVQIYPSTLVFSSPAVSTTIII  145
            CVSVP VST  +YTIPTTP  TIVSTPVQ+YPST V S+P V+TTII+
Sbjct  16   CVSVPVVSTQTVYTIPTTP--TIVSTPVQVYPSTFVVSNPTVNTTIIL  61


> PocGH01_14020600.1-p1 | transcript=PocGH01_14020600.1 | gene=PocGH01_14020600
| organism=Plasmodium_ovale_curtisi_GH01 |
gene_product=apical ring associated protein 1, putative | transcript_product=apical
ring associated protein 1, putative
| location=PocGH01_14_v1:443981-444172(-) | protein_length=63
| sequence_SO=chromosome | SO=protein_coding
Length=63

 Score = 57.0 bits (136),  Expect = 2e-11, Method: Compositional matrix adjust.
 Identities = 31/48 (65%), Positives = 36/48 (75%), Gaps = 0/48 (0%)
 Frame = +2

Query  2    CVSVPAVSTSRIYTIPTTPVSTIVSTPVQIYPSTLVFSSPAVSTTIII  145
            C+SVP +STS +Y IP T    IVSTPVQ+YPST VFSSP   TTII+
Sbjct  16   CMSVPVISTSMVYAIPATTAPAIVSTPVQVYPSTFVFSSPVQRTTIIM  63
```

**Figure 30:** The pairwise alignments between the *P. malariae* genomic DNA and the *P. relictum*/*P. ovale* annotated proteins (using PlasmoDB BLASTX online tool).

*Merozoite Surface Protein 3* (MSP3) is an important member and vaccine candidate of a multigene family expressed during the merozoite blood-stage infection of *Plasmodium* species (Beeson et al., 2016; Oeuvray et al., 1994). Screening of *P. falciparum* genome-wide expression library using Antibody Dependent Cellular Inhibition (ADCI) assay demonstrated that MSP3 is targeted by the naturally occurring host antibodies proved to be lethal for the parasite (Beeson et al., 2016; Imam et al., 2014). Since its characterization in *P. falciparum* strains (McColl et al., 1994; Oeuvray et al., 1994; Pattaradilokrat et al., 2016) studies showed that it is a 48kDa protein with a highly conserved C-terminal domain, possesses a Duffy-binding like domain and that binds to RBCs suggesting a functional role of assisting the merozoite attachment to RBCs (Beeson et al., 2016; Sakamoto et al., 2012; Singh et al., 2009). Most members of the MSP3 gene family (MSP3.1–MSP3.8) contain the *NLRNA/G* signature N-terminal peptide but it appears that 2 (MSP3.5/MSP3.6) out the 8 MSP3 proteins lack the conserved C-terminal domain and share less than 54% sequence similarity (Singh et al., 2009). Additionally, *P. vivax* and *P. knowlesi* MSP3 orthologues and paralogues have a heptad Alanine (A)-rich domain coiled-coil domain similar to a shorter coiled-coil region of the P. falciparum MSP3.1 protein (McColl et al., 1994). Even though the *P. ovale* MSP3 protein (COGENT id: POVA-CURT-01-2511)

share this N-terminal signature motif shortly after the signal peptide as illustrated in literature, statistically significant sequence similarity was observed only in *P. ovale* proteome both when searched inside the *Plasmodium* pan-genome, NR database and PlasmoDB Annotated Proteins. This observation suggests that the functional annotation is correct but it's probably an MSP3 protein fragment and thus, excluded from further analysis.

**Putative *De novo* genes**

'Orphan' or 'de novo' genes are considered those genes without detectable sequence similarity (usually BLASTP cut-off value E $<10^{-5}$ or E $< 10^{-10}$) in the genomes of other organisms and are genes which do not encode any previously-identified protein domains. However, increasing evidence from genomic and transcriptomic sequencing show that these genes might have a narrow phylogenetic distribution (i.e. homologous 'orphan' can be present in closely related species but are not present in more distantly related species or in other genera) (Khalturin et al., 2009; Wilson et al., 2005).

The initial BLASTP set of putative *de novo* genes (i.e. before the elimination of partial sequences) consisted of 1011 proteins (including the 9 proteins with no BLASTP hits; Table 14,Table 15). Combining the results from the tBLASTN analysis only 638 unique proteins were left as putative *de novo* and after the elimination of all partial sequences we reduced this number to 96 putative *de novo* proteins (**Table 18**, Table 19). Each of these proteins were further inspected using the integrated BLAST suite of tools of PlasmoDB (C. Aurrecoechea et al., 2009) and Pfam protein sequence analysis online tool (Finn et al., 2016, 2006).

Our in-depth analysis revealed that for 90 putative *de novo* protein sequences (apart from a *P. vivax* protein with transcriptomic data), no statistically significant sequence similarity was detected in the genomes of other *Plasmodium* species and the Pfam analysis did not detect any previously-identified protein domains. In PlasmoDB gene record of the *P. vivax* protein (COGENT id: PVIV-Sal1-01-902), there are transcriptomic data of its expression during the 48hr intraerythrocytic cycle using microarray and RNA-seq analysis (Bozdech et al., 2008; Zhu et al., 2016). Both studies illustrated a strain specific transcriptional regulation for a subset of *P. vivax* proteins including this protein sequence suggesting an important functional role in *P. vivax* life cycle (Bozdech et al., 2008; Zhu et al., 2016).

Furthermore, for some of these de novo proteins we observed weak hits (i.e. the hits e-value was below our e-value cutoff: e-value > $1e^{-6}$) in other closely related *Plasmodium* species genomes but either the aligned hits were due to CBRs and considered FP or the hits were in an unannotated genomic locus suggesting that these genes are probably unique due to gene prediction missed-annotation artefact. For the remaining 6 orphan proteins, Pfam analysis detected statistical significant match to a previously-defined protein domain (**Table 18**, **Table 19**).

An interesting case of this category is a *P. fragile* protein (COGENT id: PFRA-NILG-01-1600) which was set as unique (only self-hits) due to small length (**Table 19**). This unusually short protein of 35 amino acid residues has significant sequence similarity with the Ost4 subunit of the oligosaccharyltransferase (OST) complex (Pfam e-value: $7.1e^{-5}$) found in a variety of species from fungi to humans but supposedly absent from *Plasmodium* (Kelleher and Gilmore, 2006; Lombard, 2016). This finding initiated the exhaustive search for subunits of the OST complex in *Plasmodium spp.* and in select protist species - the alveolate *Cryptosporidium parvum* (phylum: Apicomplexa), the excavate *Trichomonas vaginalis* (phylum: Metamonada), and *Entamoeba histolytica* (from the Amoebozoa supergroup) - as sample species of this diverse group of unicellular eukaryotes. The results of this analysis are presented in Chapter3.3.

Accordingly, is intriguing why the three *P. malariae* proteins annotated as "hypothetical protein" clustered as de novo (**Table 19**), even though, we observed statistically significant similarity to the Secreted Polymorphic Antigen Pfam protein domain (SPAM; PF07133.10) associated with MSPs. During blood-stage infection, *Plasmodium* species, while in merozoite form, secretes surface antigens that undergo excessive processing pre- and post-translation and assist in the invasion of RBCs (Beeson et al., 2016; Guha-Niyogi et al., 2001). MSPs have been targeted as novel vaccines candidates as an effort to successfully eradicate malaria (Beeson et al., 2016; Crosnier et al., 2013; Oeuvray et al., 1994; Singh et al., 2009). Thus, is it due to heavily biased sequences in which filtering CBRs results in failure of the comparative genomics analysis pipeline or the sequences are so divergent due to deletions and amino acid substitutions that leave few traces of sequence similarity or gene prediction artefacts?

All sequence comparisons we performed (pan-genome, NR/NT and in PlasmoDB) using *masked* protein sequences confirmed their de novo status as the only statistically significant sequence similarity (e-value cut-off: $1e^{-6}$) was itself or in same-species sequences (i.e. paralogous sequences). However, apart from PMAL-UG01-01-1341, the other two proteins showed statistically significant similarity to the SPAM protein domain which, according to the literature is associated with MSPs (McColl et al., 1994; Mulhern et al., 1995). This protein family consists of an unusual alanine heptad-repeat domain and a conserved hydrophilic C-terminal domain (McColl et al., 1994; Mulhern et al., 1995). Additionally, it consists one of the first examples that repetitive regions have well-defined structural and functional elements (McColl et al., 1994).

Even though our comparative genomics analysis set PMAL-UG01-01-1341 protein as de novo and Pfam protein sequence analysis did not detect any known protein domain, in PlasmoDBs' gene record is set in an orthologous group (OG5_126560) with MSP3 from other *Plasmodium* species. Thus, we first wanted to check if this sequence contains Asn-X-Ser/Thr sequons for N-glycosylation, which a typical for a subtype of MSP3s and secondly, confirm that it is not a gene prediction artefact. First, using NetNGlyc server (Gupta et al., 2004) we confirmed that this protein contains multiple sequons at the N-terminal as shown in **Figure 31**. Then, we downloaded all proteins from UniProt/TrEMBL annotated as MSP3 (search term: "merozoite surface protein 3" organism: *Plasmodium*) in FASTA format, a dataset composed of 1180 protein sequences.

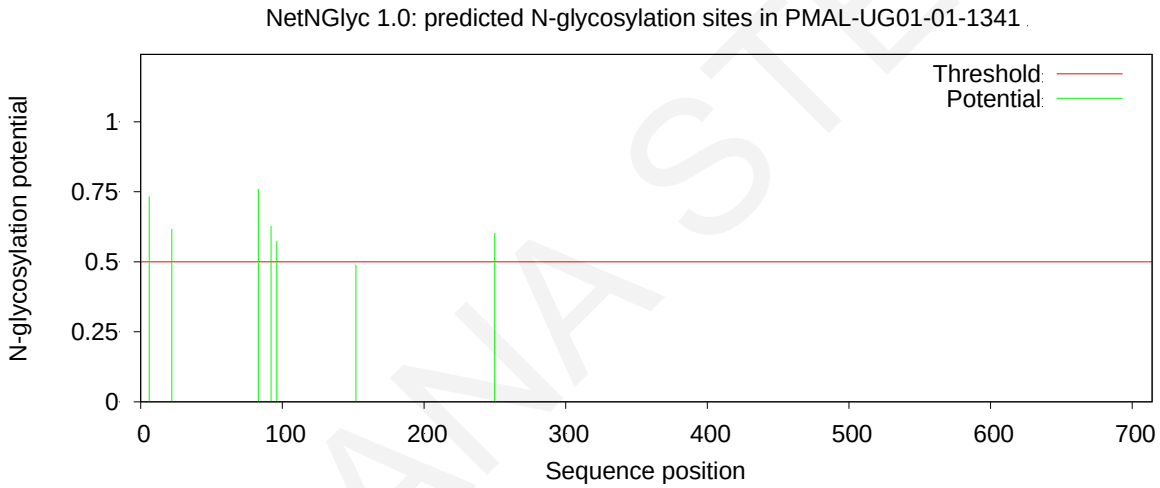



**Figure 31:** NetNGlyc output for the PMAL-UG01-01-1341 protein sequence which confirms the presence of the Asn-X-Ser/Thr sequon at the N-terminal.

We performed BLASTP pairwise alignments using the online NCBI BLASTP tool (with default parameters) and selected for download the pairwise alignments only for those showing statistically significant local sequence similarity (e-value cutoff: $1e^{-6}$) leading to a total three sequences, two from *P. malariae* (34% similarity) and one from *P. gallinaceum* (35% similarity). The pairwise alignments were then subjected into mView (Brown et al., 1998; McWilliam et al., 2013) for graphical representation of the local alignments between these sequences (**Figure 32**). Reviewing the MSA, it's evident that the observed local sequence similarity of the PMAL-UG01-1341 with the other three MSP3 proteins is at the N-terminal where the MSP3 signature peptide is located but the sequences are highly divergent that very few traces left of a possible

common ancestor sequence. Furthermore, we excluded the possibility of gene prediction artefact by extracting a 5000bp genomic DNA sequence for this protein (using the Sequence Retrieval tool from PlasmoDB website) and performed gene prediction using FGENESH online (Solovyev et al., 2006). Protein sequence comparison of the predicted protein sequence (659 amino acids) gives as best statistically significant match (e-value: 0.0) back the PMAL-UG01-01-1341 protein sequence. These results suggest that we correctly detected this protein as *de novo*.

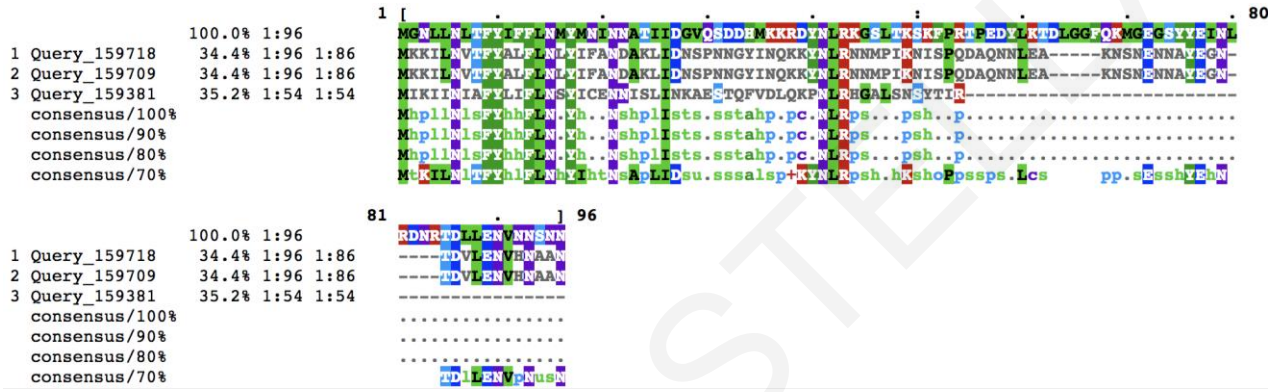



**Figure 32:** A graphical representation of the pairwise alignments between PMAL-UG01-01-1341 and the three MSP3 proteins retrieved from UniProt/TrEMBL.

**Table 18:** A list of the *Plasmodium* species putative Orphan proteins.

| A/A | COGENT ID | PlasmoDB ID | Protein Length | Annotation | CBRs | %CBRs | %GC | %AT | ND | QIPP_wCBRs | QIPP_woCBRs | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1.** | PKNO-H-01-4915 | PKNH_1441900 | 305 | Plasmodium exported protein, unknown function | K, K, E | 19,02 | 37,36 | 62,64 | 0,9 | 0,63 | 0,51 | Probably not annotated in other Plasmodium species |
| **2.** | PMAL-UG01-01-1171 | PmUG01_05040600 | 86 | hypothetical protein | N | 8.14 | 34.48 | 65.52 | 0.5 | 0.41 | 0.22 | Species specific |
| **3.** | PMAL-UG01-01-1188 | PmUG01_05042300 | 78 | hypothetical protein | N/A | 0 | 27,85 | 72,15 | 0 | 0,53 | 0,38 | Strain specific |
| **4.** | PMAL-UG01-01-1338 | PmUG01_06022400 | 220 | hypothetical protein | E | 17,73 | 34,24 | 65,76 | 0 | 0,35 | 0,14 | Probably not annotated in other Plasmodium species |
| **5.** | PMAL-UG01-01-1364 | PmUG01_06025000 | 124 | hypothetical protein | K | 16,13 | 31,73 | 68,27 | 0,4 | 0,43 | 0,25 | Strain specific |
| **6.** | PMAL-UG01-01-1422 | PmUG01_07013700 | 226 | hypothetical protein | E | 6.19 | 37 | 63 | 0 | 0.34 | 0.12 | Strain specific |
| **7.** | PMAL-UG01-01-1760 | PmUG01_07047800 | 461 | hypothetical protein | E, D, K | 23,64 | 30,88 | 69,12 | 0,2 | 0,45 | 0,27 | PlasmoDB OG5_203186 |
| **8.** | PMAL-UG01-01-259 | PmUG01_02011700 | 98 | hypothetical protein | N/A | 0 | 28,28 | 71,72 | 0,00 | 0,56 | 0,41 | Probably not annotated in other Plasmodium species |
| **9.** | PMAL-UG01-01-3164 | PmUG01_10046500 | 355 | hypothetical protein | E, N, D | 20 | 37,83 | 62,17 | 0,6 | 0,45 | 0,27 | Strain specific |
| **10.** | PMAL- | PmUG01_11015500 | 125 | hypothetical | N/A | 0 | 32,54 | 67,46 | 0,2 | 0,39 | 0,18 | Strain |

| A/A | COGENT ID | PlasmoDB ID | Protein Length | Annotation | CBRs | %CBRs | %GC | %AT | ND | QIPP_wCBRs | QIPP_woCBRs | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | UG01-01-3313 | | | protein | | | | | | | | specific |
| **11.** | PMAL-UG01-01-3314 | PmUG01_11015600 | 306 | Plasmodium exported protein, unknown function | N/A | 0 | 23,24 | 76,76 | 0,3 | 0,42 | 0,23 | Strain specific |
| **12.** | PMAL-UG01-01-35 | PmUG01_01013500 | 267 | hypothetical protein | N | 8,61 | 23,51 | 76,49 | 0 | 0,38 | 0,18 | Strain specific |
| **13.** | PMAL-UG01-01-3801 | PmUG01_12011800 | 85 | hypothetical protein | E | 20 | 34,11 | 65,89 | 0,5 | 0,42 | 0,22 | Strain specific |
| **14.** | PMAL-UG01-01-4477 | PmUG01_12080700 | 213 | conserved Plasmodium protein, unknown function | T, K | 17,84 | 31,93 | 68,07 | 0,7 | 0,5 | 0,34 | Strain specific |
| **15.** | PMAL-UG01-01-4509 | PmUG01_13011500 | 225 | hypothetical protein | I | 4 | 26,55 | 73,45 | 0,4 | 0,54 | 0,38 | Strain specific |
| **16.** | PMAL-UG01-01-5129 | PmUG01_14014200 | 118 | hypothetical protein | N/A | 0 | 26,89 | 73,11 | 0 | 0,49 | 0,32 | Strain specific |
| **17.** | PMAL-UG01-01-6132 | PmUG01_00051300 | 87 | hypothetical protein | E | 8,05 | 29,54 | 70,46 | 0,5 | 0,5 | 0,34 | Strain specific |
| **18.** | PMAL-UG01-01-6123 | PmUG01_00050400 | 429 | conserved Plasmodium protein, unknown function | N, K | 26.11 | 23.58 | 76.42 | 0.9 | 0.56 | 0.42 | Strain specific |
| **19.** | POVA-CURT-01-1069 | PocGH01_06028200 | 460 | hypothetical protein | E, C, N | 24.13 | 41.58 | 58.42 | 0.4 | 0.45 | 0.27 | Species specific |

| A/A | COGENT ID | PlasmoDB ID | Protein Length | Annotation | CBRs | %CBRs | %GC | %AT | ND | QIPP_wCBRs | QIPP_woCBRs | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **20.** | POVA-CURT-01-1425 | PocGH01_07043500 | 335 | hypothetical protein | E, D, K, N | 25.07 | 32.04 | 67.96 | 0.3 | 0.53 | 0.38 | Species specific |
| **21.** | POVA-CURT-01-1073 | PocGH01_06028600 | 613 | hypothetical protein | E, Q, K, N | 29,69 | 41,97 | 58,03 | 0,4 | 0,47 | 0,3 | Strain specific |
| **22.** | POVA-CURT-01-1791 | PocGH01_08046200 | 85 | hypothetical protein | N/A | 0 | 25,19 | 74,81 | 0,2 | 0,4 | 0,2 | Strain specific |
| **23.** | POVA-CURT-01-1788 | PocGH01_08045900 | 323 | hypothetical protein | E | 8.98 | 36.21 | 63.79 | 0.5 | 0.49 | 0.32 | Species specific |
| **24.** | POVA-CURT-01-2541 | PocGH01_11010200 | 168 | hypothetical protein | N/A | 0 | 27.42 | 72.58 | 0.4 | 0.49 | 0.33 | Species specific |
| **25.** | POVA-CURT-01-3651 | PocGH01_12078500 | 395 | conserved Plasmodium protein, unknown function | T, N, G, E | 24,3 | 36,62 | 63,38 | 0,5 | 0,49 | 0,33 | Strain specific |
| **26.** | POVA-CURT-01-3661 | PocGH01_12079500 | 102 | hypothetical protein | E | 8.82 | 35.28 | 64.72 | 0.1 | 0.41 | 0.21 | Species specific |
| **27.** | POVA-CURT-01-3675 | PocGH01_13011200 | 87 | hypothetical protein | N/A | 0 | 18,94 | 81,06 | 0,2 | 0,37 | 0,16 | Strain specific |
| **28.** | POVA-CURT-01-4701 | PocGH01_14068000 | 149 | hypothetical protein | N/A | 0 | 28,22 | 71,78 | 0,8 | 0,6 | 0,46 | Strain specific |
| **29.** | POVA-CURT-01-4702 | PocGH01_14068100 | 74 | hypothetical protein | N/A | 0 | 33,78 | 66,22 | 0,8 | 0,57 | 0,42 | Strain specific |
| **30.** | POVA-CURT-01-4878 | PocGH01_00077600 | 128 | hypothetical protein | Y | 4.69 | 29.36 | 70.64 | 0 | 0.49 | 0.32 | Species specific |

| A/A | COGENT ID | PlasmoDB ID | Protein Length | Annotation | CBRs | %CBRs | %GC | %AT | ND | QIPP_wCBRs | QIPP_woCBRs | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **31.** | POVA-CURT-01-5242 | PocGH01_00025100 | 147 | hypothetical protein | N/A | 0 | 25.72 | 74.28 | 0 | 0.39 | 0.19 | Species specific |
| **32.** | POVA-CURT-01-5244 | PocGH01_00109400 | 92 | hypothetical protein | N | 11.96 | 27.86 | 72.14 | 0 | 0.43 | 0.25 | Species specific |
| **33.** | POVA-CURT-01-5374 | PocGH01_00120800 | 207 | hypothetical protein | N/A | 0 | 31.4 | 68.6 | 0 | 0.58 | 0.44 | Species specific |
| **34.** | POVA-CURT-01-5390 | PocGH01_00011500 | 320 | hypothetical protein | N, E | 8,44 | 33,74 | 66,26 | 0 | 0,45 | 0,27 | Strain specific |
| **35.** | POVA-CURT-01-5554 | PocGH01_00029200 | 88 | hypothetical protein | K | 15.91 | 28.16 | 71.84 | 0 | 0.44 | 0.26 | Species specific |
| **36.** | POVA-CURT-01-5722 | PocGH01_00147500 | 86 | hypothetical protein | N/A | 0 | 30.68 | 69.32 | 0 | 0.54 | 0.38 | Species specific |
| **37.** | POVA-CURT-01-6231 | PocGH01_00184800 | 373 | Plasmodium exported protein, unknown function | K | 17.16 | 32.47 | 67.53 | 0 | 0.48 | 0.31 | Species specific |
| **38.** | POVA-CURT-01-6353 | PocGH01_00194000 | 74 | hypothetical protein | N/A | 0 | 28,44 | 71,56 | 0 | 0,45 | 0,27 | Strain specific |
| **39.** | POVA-CURT-01-6573 | PocGH01_00211500 | 92 | hypothetical protein | N, G | 39,13 | 34,3 | 65,7 | 0,1 | 0,42 | 0,23 | Strain specific |
| **40.** | POVA-CURT-01-6810 | PocGH01_00232200 | 129 | hypothetical protein | N/A | 0 | 28,9 | 71,1 | 0 | 0,47 | 0,3 | Strain specific |
| **41.** | POVA-CURT-01-6832 | PocGH01_00234400 | 84 | hypothetical protein | N/A | 0 | 26,69 | 73,31 | 0 | 0,4 | 0,2 | Strain specific |

| A/A | COGENT ID | PlasmoDB ID | Protein Length | Annotation | CBRs | %CBRs | %GC | %AT | ND | QIPP_wCBRs | QIPP_woCBRs | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **42.** | POVA-CURT-01-6833 | PocGH01_00234500 | 219 | hypothetical protein | N/A | 0 | 29.04 | 70.96 | 0 | 0.49 | 0.32 | Species specific |
| **43.** | POVA-CURT-01-6834 | PocGH01_00234600 | 108 | hypothetical protein | N | 16,67 | 28,79 | 71,21 | 0 | 0,46 | 0,29 | Strain specific |
| **44.** | POVA-CURT-01-6859 | PocGH01_00237100 | 99 | hypothetical protein | N/A | 0 | 28,74 | 71,26 | 0 | 0,46 | 0,28 | Strain specific |
| **45.** | POVA-CURT-01-7004 | PocGH01_00060800 | 148 | hypothetical protein | K, N | 28.38 | 28.38 | 71.62 | 0 | 0.45 | 0.28 | Species specific |
| **46.** | POVA-CURT-01-7051 | PocGH01_00064700 | 89 | hypothetical protein | N/A | 0 | 26.67 | 73.33 | 0 | 0.4 | 0.2 | Species specific |
| **47.** | POVA-CURT-01-7098 | PocGH01_00068500 | 316 | Plasmodium exported protein, unknown function | E, D, K | 29,43 | 27,51 | 72,49 | 0 | 0,45 | 0,27 | Strain specific |
| **48.** | POVA-CURT-01-7101 | PocGH01_00068800 | 137 | hypothetical protein | N/A | 0 | 29.46 | 70.54 | 0 | 0.49 | 0.32 | Species specific |
| **49.** | POVA-CURT-01-902 | PocGH01_06011500 | 234 | hypothetical protein | D, N, E | 23.93 | 31.91 | 68.09 | 0 | 0.48 | 0.31 | Species specific |
| **50.** | PVIN-PECR-01-2527 | YYG_02558 | 162 | hypothetical protein | E | 8.64 | 34.36 | 65.64 | 0 | 0.33 | 0.11 | Species specific |
| **51.** | PVIV-Sal1-01-902 | PVX_089440 | 50 | hypothetical protein | K | 28 | 40,52 | 59,48 | 0 | 0,44 | 0,26 | (Bozdech et al., 2008) |

**Table 19:** A list of *Plasmodium* Orphan proteins with statistically significant sequence similarity to other *Plasmodium* species proteins or with known Pfam protein family domains.

| Protein Family | COGENT ID | PlasmoDB | Length | Annotation | CBRs | Evidence | Notes | Reference |
|---|---|---|---|---|---|---|---|---|
| **Ost4 subunit** | PFRA-NILG-01-1600 | AK88_01616 | 35 | hypothetical protein | N/A | Pfam e-value: $4.3e^{-09}$ | Not annotated in other *Plasmodium* genomes | (Tamana and Promponas, 2018) |
| **Merozoite surface protein** | PMAL-UG01-01-1341 | PmUG01_06022700 | 714 | hypothetical protein | E, K | Signature peptide of MSP3<br>Significant sequence similarity to *P. malariae* MSP3 ($6e^{-07}$)<br>PlasmoDB OG5_126560 | Missed due to heavy masking | (Beeson et al., 2016; McColl et al., 1994; Mulhern et al., 1995) |
| | PMAL-UG01-01-3162 | PmUG01_10046300 | 605 | hypothetical protein | E, S, D | Pfam e-value: $1.1e^{-13}$<br>Significant hits to P. reichenowi (e-value: $6e^{-07}$), P. falciparum ($2e^{-06}$), P. coatneyi (e-value: $2e^{-04}$), P gallinaceum (e-value: $9e^{-04}$) | Missed due to heavy masking | |
| | PMAL-UG01-01-3165 | PmUG01_10046600 | 405 | hypothetical protein | E, N | Pfam e-value: $2e^{-12}$<br>Significant hits to P. reichenowi (e-value: $8e^{-04}$), P. falciparum IT ($1e^{-4}$), P gallinaceum (e-value: $3e^{-04}$) | Missed due to heavy masking | |
| | PMAL-UG01-01-3167 | PmUG01_10046800 | 406 | hypothetical protein | E, D, N | Pfam e-value: $3.7e^{-8}$<br>TBLASTN significant hit with P. knowlesi (e-value: $8e^{-04}$) | Missed due to heavy masking | |
| | PMAL-UG01-01-3171 | PmUG01_10047200 | 423 | hypothetical protein | E | Pfam e-value: $3.5e^{-7}$ | Missed due to heavy masking | |

| Protein Family | COGENT ID | PlasmoDB | Length | Annotation | CBRs | Evidence | Notes | Reference |
|---|---|---|---|---|---|---|---|---|
| **PIR protein** | PMAL-UG01-01-201 | PmUG01_01031100 | 99 | hypothetical protein | N/A | PlasmoDB OG5_127519 Significant hits with P. vivax (e-value: $7e^{-6}$) | Probably partial/fragment | (Cunningham et al., 2010; del Portillo et al., 2001; Janssen et al., 2002) |
| | PMAL-UG01-01-2830 | PmUG01_10012400 | 238 | hypothetical protein | N/A | Significant hits with *P. vivax* (e-value: $2e^{-06}$) and *P. ovale* (e-value: $8e^{-06}$) | Probably partial/fragment | |
| | PMAL-UG01-01-3310 | PmUG01_11015200 | 76 | hypothetical protein | N/A | BLASTX (genomic DNA) significant hits with P. vivax (e-value: $2e^{-04}$), P. ovale (e-value: $7e^{-05}$) and P. cynomolgi (e-value: $6e^{-05}$) | Probably partial/fragment | |
| | POVA-CURT-01-5485 | PocGH01_00128500 | 130 | hypothetical protein | N/A | TBLASTN significant hits to *P. ovale* PIR proteins Pfam e-value: $6.0e^{-07}$ | Probably partial/fragment | |
| | POVA-CURT-01-6368 | PocGH01_00044500 | 148 | hypothetical protein | N/A | BLASTP significant hits to P. ovale PIR proteins Pfam match to AAA+ domain (e-value: $2.9e^{-07}$) | Probably partial/fragment | |
| | POVA-CURT-01-6375 | PocGH01_00195900 | 107 | hypothetical protein | N/A | BLASTP significant hits to P. ovale PIR proteins | Probably partial/fragment | |
| | POVA-CURT-01-6433 | PocGH01_00200100 | 101 | hypothetical protein | N/A | Weak matched to P. ovale PIR proteins | Probably partial/fragment | |
| | POVA-CURT-01-6945 | PocGH01_00055000 | 201 | hypothetical protein | E | BLASTP significant hits to P. ovale PIR proteins | Probably partial/fragment | |
| | POVA-CURT-01-7002 | PocGH01_00060600 | 124 | hypothetical protein | Y | BLASTP significant hits to P. ovale PIR proteins | Probably partial/fragment | |
| **Early** | PMAL- | PmUG01_11061000 | 168 | hypothetical | E | Pfam e-value: $9.1e^{-07}$ | Strain specific | (Favaloro et |

| Protein Family | COGENT ID | PlasmoDB | Length | Annotation | CBRs | Evidence | Notes | Reference |
|---|---|---|---|---|---|---|---|---|
| **Transcribed Membrane Protein (ETRAMP)** | UG01-01-3760 | | | protein | | | | al., 1993; Spielmann and Beck, 2000) |
| **Small Exported Protein (SEMP1)** | PMAL-UG01-01-39 | PmUG01_01013900 | 69 | hypothetical protein | N/A | Significant hits with *P. ovale* (e-value: $6e^{-06}$), *P. vivax* (e-value: $2e^{-05}$) and *P. cynomolgi* (e-value: $3e^{-04}$) | Missed due to small length | (Dietz et al., 2014) |
| **KS1-precursor** | PMAL-UG01-01-4431 | PmUG01_12075900 | 239 | conserved Plasmodium protein, unknown function | D, E, K | Significant hits with *P. vivax* KS1-precursor protein (e-value: $8e^{-14}$) | Missed due to heavy masking TRG | - |
| **STP1 protein** | PMAL-UG01-01-5133 | PmUG01_14014600 | 67 | hypothetical protein | N/A | Significant hits to *P. malariae* STP1 proteins/psuedogenes | Probably pseudogene Species specific | (Rutledge et al., 2017) |
| | PMAL-UG01-01-6149 | PmUG01_00052900 | 188 | hypothetical protein | N, S | Significant hits to *P. malariae* STP1 proteins/psuedogenes | Probably pseudogene Species specific | |
| **Unknown function** | PMAL-UG01-01-2136 | PmUG01_08043300 | 197 | conserved Plasmodium protein, unknown function | N | Significant hits with P. fragile (e-value: $2e^{-04}$) and P. ovale (e-value: $2e^{-06}$) | Missed to heavy masking and strict e-value TRG | - |
| | PMAL-UG01-01-2213 | PmUG01_08051500 | 332 | hypothetical protein | D, S | Significant hits with P. fragile (e-value: $7e^{-04}$) | Missed to heavy masking and strict e-value TRG | |
| | PMAL-UG01-01-2522 | PmUG01_09028500 | 151 | conserved protein, unknown | K | PlasmoDB OG5_131402 Pfam match to DUF1764 ($3.1e^{-09}$) | Missed to heavy masking and strict e-value | |

| Protein Family | COGENT ID | PlasmoDB | Length | Annotation | CBRs | Evidence | Notes | Reference |
|---|---|---|---|---|---|---|---|---|
| | | | | function | | Significant hits to all Plasmodium Species | Not unique | |
| | PMAL-UG01-01-4351 | PmUG01_12067800 | 69 | conserved Plasmodium protein, unknown function | K | Significant hits with P. gaboni, P. falciparum (e-value: $5e^{-12}$), P. reichenowi (e-value: $5e^{-12}$), P. gaboni (e-value: $2e^{-12}$), P. gallinaceum (e-value: $4e^{-08}$) and P. relictum (e-value: $9e^{-05}$) | Missed due to heavy masking TRG | |
| | PMAL-UG01-01-4509 | PmUG01_13011500 | 225 | hypothetical protein | I | Pfam matches to Peptidase_C97 and DUF2207 | Strain specific | - |
| | PMAL-UG01-01-4528 | PmUG01_13013400 | 155 | conserved Plasmodium protein, unknown function | N, E, K | Significant hits with P. fragile (e-value: $6e^{-08}$), P. gallinaceum (e-value: $4e^{-16}$), P. relictum (e-value: $1e^{-14}$), P. ovale (e-value: $1e^{-11}$), P. vivax (e-value: $5e^{-11}$), P. knowlesi (e-value: $3e^{-09}$), P. inui (e-value: $7e^{-08}$), P. chabaudi (e-value: $7e^{-04}$), P. coatneyi (e-value: $5e^{-08}$) and P. vinckei petteri (e-value: $7e^{-04}$) PlasmoDB OG5_167070 | Missed due to heavy masking TRG | - |
| | PMAL-UG01-01-5147 | PmUG01_14015900 | 92 | hypothetical protein | N/A | TBLASTN significant hit with P. ovale (e-value: $9e^{-05}$) | Missed due to strict e-value TRG | - |
| | POVA-CURT- | PocGH01_12065700 | 88 | conserved Plasmodium | K | Significant hits with P. falciparum (e-value: $3e^{-}$ | Missed due to heavy masking | - |

| Protein Family | COGENT ID | PlasmoDB | Length | Annotation | CBRs | Evidence | Notes | Reference |
|---|---|---|---|---|---|---|---|---|
| | 01-3526 | | | protein, unknown function | | $^{09}$), P. reichenowi (e-value: 3e$^{-09}$), P. gaboni (e-value: 7e$^{-09}$), P. knowlesi (e-value: 1e$^{-07}$), P. gallinaceum (e-value: 2e$^{-06}$) and P. malariae (e-value: 1e$^{-04}$) | TRG | |

***Plasmodium* Taxonomically Restricted Genes**

In this section, we will review the initially putative unique genes restricted in a specific *Plasmodium* lineage based on the *Plasmodium* core genome phylogenetic tree and primary parasite host (**Figure** 18; **Table 20**). Studying of lineage-specific genes could shed light on the evolutionary mechanisms shaping the diversity and pathogenicity of *Plasmodium* species. In literature, TRGs are considered of utmost interest since their origination does not always follow the classical model of evolution, e.g. duplication, rearrangement, and mutation of genes inherited from a common ancestor (Khalturin et al., 2009; Tautz and Domazet-Lošo, 2011; Wilson et al., 2005; Zhou et al., 2008) but rather, arise directly from previously noncoding regions (Verster et al., 2017). In general, TRGs are characterized by short, repetitive or unusual A+T content sequences (Tautz and Domazet-Lošo, 2011) or contain no previously-determined protein domain and have detectable homology only in genomes of highly related species (Verster et al., 2017).

After careful analysis, we finalized a set of 25 TRG protein sequences where statistically significant sequence similarity was observed in a subset of *Plasmodium* species (**Table 20**). Almost all these proteins (except from some *P. vinckei* proteins with BLASTP hits in *P. chabaudi*) were classified as Strain/Species specific at the initial sequence comparison step (**Figure 17**; BLASTP e-value: $1e^{-06}$) since the observed statistical significant similarity was in self-strain/species. However, for most of these proteins the search for genes encoding similar protein as the query (**Figure 17**; tBLASTN e-value: $1e^{-6}$) detected statistically significant sequence similarity in closely related *Plasmodium* species (**Table 20**) and in two cases (COGENT ids: POVA-CURT-01-6890 and PVIN-PECR-01-2106) in species from other genera such as *Salmon salar* (phylum: Chordata), the excavate *Leishmania braziliensis* (phylum: Euglenozoa) and the bacteria *Cupriavidus sp. USMAA2-4* (phylum: Proteobacteria). For all these protein sequences we performed BLASTP/tBLASTN searches against PlasmoDBs' Annotated Proteins/Genome as to cross-reference the NCBI *Plasmodium* BLASTP/tBLASTN results and eliminate possible spurious hits. For the three cases with hits to other genera, we further inspected the tBLASTN hits by extracting and visualizing the pairwise alignments using the standalone version of mView (Brown et al., 1998; McWilliam et al., 2013). Specifically, for each of these two proteins, we formatted its tBLASTN hits file as a BLASTN database and re-run tBLASTN using the respective protein as the query sequence. The pairwise alignments were set as the required input file to mview, for the

reconstruction of MSAs. An example of the mView output and the pairwise constructed MSA can be seen in **Figure 33**.

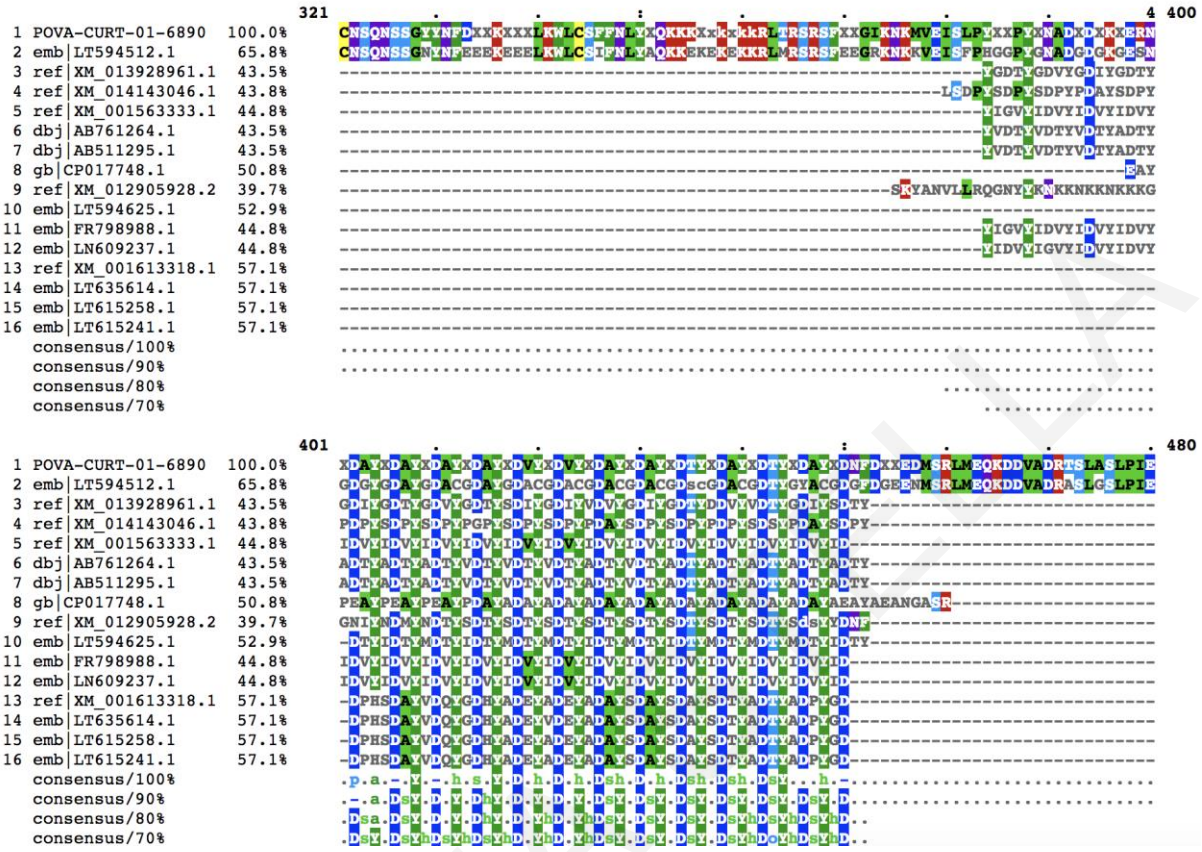



**Figure 33:** A graphical representation of the MSA constructed by the POVA-CURT-01-6890 pairwise alignments.

Sadly, all tBLASTN hits from other genera were discarded as FP since the initial alignments were due to certain amino acid repeats (**Figure 33**). Thus, these 25 proteins were classified as TRG after we confirmed that there are genes, in closely related Plasmodium, encoding proteins with statistically significant sequence similarity to these sequences (e-value cutoff: $1e^{-06}$).

**Table 20:** Summary table for the 26 TRG proteins based on the *Plasmodium* core genome phylogenetic tree. The *Clade* column signifies the *Plasmodium* species clade while, the *TRG Clade* signifies the clades which statistically significant sequence similarity was observed for the respective TRG protein.

| Species | #TRG proteins | TRG Clade |
|---|---|---|
| ***P. cynomolgi* B** | PCYN-B-01-1628 \| PCYN-B-01-3808 | P. falciparum/P. cynomolgi |
| ***P. knowlesi* H** | PKNO-H-01-340 | P. coatney/P.vivax |
| ***P. ovale curtisi*** | POVA-CURT-01-1071 \| POVA-CURT-01-5767 \| POVA-CURT-01-6105 | P. malariae/P. ovale |
| | POVA-CURT-01-1068 \| POVA-CURT-01-5887 \| POVA-CURT-01-6890 | Human-Simians clade |
| ***P. vinckei* petteri** | PVIN-PECR-01-1308 \| PVIN-PECR-01-1137 \| PVIN-PECR-01-21<br>PVIN-PECR-01-2618 \| PVIN-PECR-01-2871<br>PVIN-PECR-01-318 \| PVIN-PECR-01-488 | *Rodent's* clade |
| | PVIN-PECR-01-2072 \| PVIN-PECR-01-2106 \| PVIN-PECR-01-3810 | P. vinckei/P. chabaudi |
| ***P. vinckei* vinckei** | PVIN-VINK-01-268 \| PVIN-VINK-01-4429 | Rodent's clade |
| | PVIN-VINK-01-4373 | P. yoelii/P. vinckei |
| ***P. vivax* P01** | PVIV-P01-01-2279 \| PVIV-P01-01-2748 \| PVIV-P01-01-3575 \| PVIV-P01-01-6106 | P. vivax/P. cynomolgi |

The Bayesian-inferred *Plasmodium* core genome phylogenetic tree (**Figure** 18) classifies the *Plasmodium* species into 5 distinct lineages: the *rodent-infecting* parasites (*P. yoelii*, *P. berghei*, *P. chabaudi* and *P. vinckei*), the *primate-/avian-infecting* parasites (*P. falciparum* and *P. reichenowi*, *P. gallinaceum* and *P. relictum*), *human-infecting* parasites (*P. malariae* and *P. ovale*) and the *simian-infecting* parasites (*P. coatneyi*, *P. knowlesi*, *P. fragile*, *P. cynomolgi*, *P. inui* and *P. vivax*). Interestingly, we observed that TRG proteins (**Figure 34**) that belong in *Plasmodium* species that infect rodents (12 proteins from *P. vinckei*; clade P) or primates (2 proteins from *P. cynomolgi*, 1 protein from *P. knowlesi* and 4 proteins from *P. vivax*; clade Q) are restricted at their respective clade as opposed to those proteins that belong to humans' parasite (all proteins from *P. ovale*; clade T). Based on the Presence/Absence heat-map we constructed using R's package pheatmap (MRAN, 2018) we observed that only two out of the six *P. ovale* proteins are restricted in human's clade while, for all proteins from rodents and simians-infecting parasites homologous sequences were observed only in their respective clade (**Figure 34**).

This observation suggests that these proteins may play an important role in the parasite adaptation and could have specific functions required only by the molecular biology of these species. In contrast, the *P. ovale* TRG proteins that share a more generic TRG profile with homologous sequences in both from rodent and primate clades, may reflect evolutionary events shaping its pathogenicity. It's worth mentioning that almost all TRG proteins are annotated either as "hypothetical protein" or as "conserved *Plasmodium* protein, unknown function" suggesting that improving functional annotation of the parasite's genomes could be crucial in rational design of highly effective drug/vaccine candidates or for optimizing diagnostic tools.

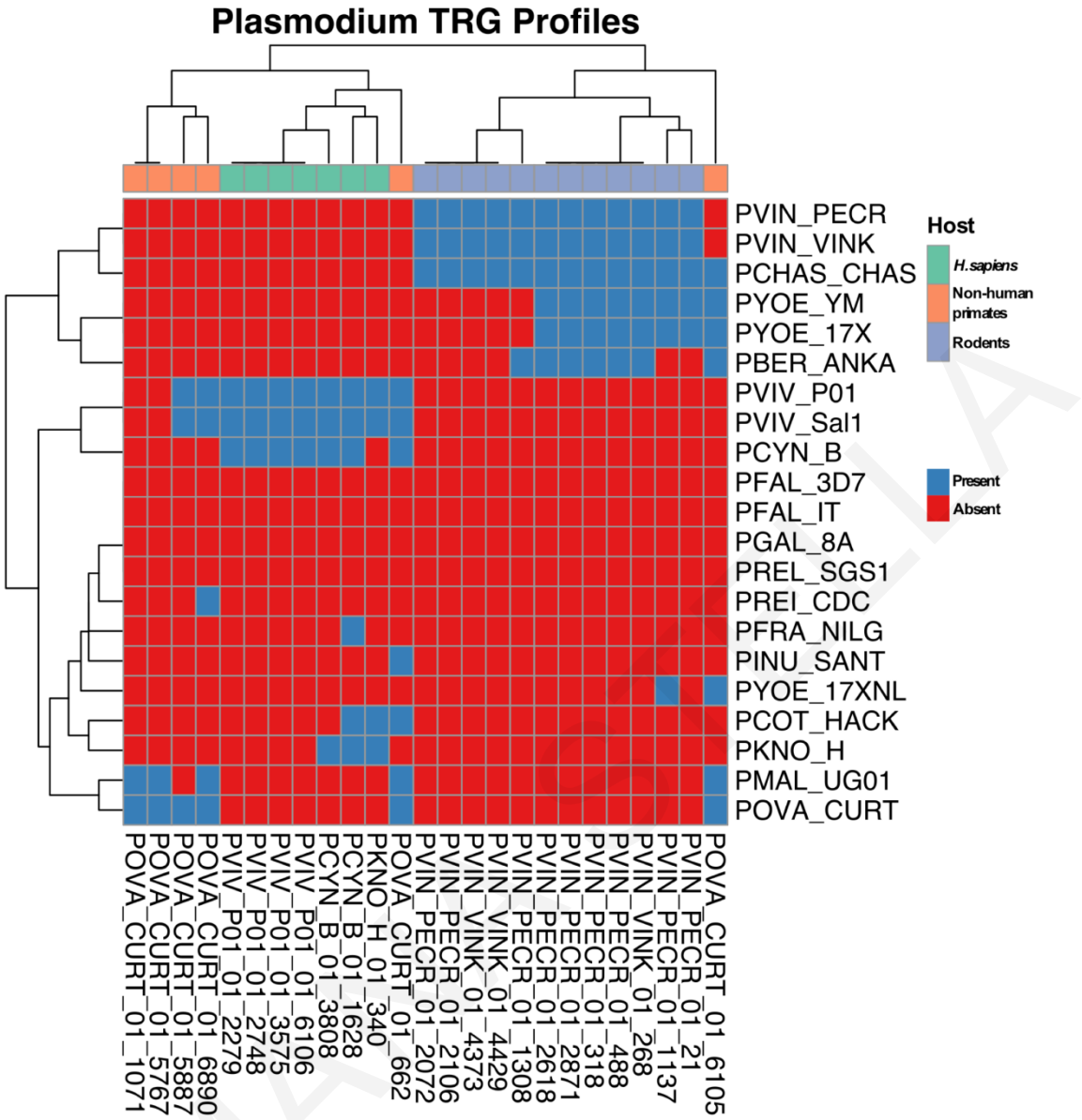



**Figure 34:** Presence/Absence heat-map phylogenetic profile of all TRG proteins. **Blue:** present genes and **Red:** absent genes from a particular *Plasmodium* species. Heat-map was constructed using R's studio package pheatmap (MRAN, 2018; RStudio Team, 2015).

### Possibly Contaminated Sequences

A contaminated sequence is one that does not actually represent the genetic information from the biological source organism/organelle because it contains one or more sequence segments of foreign origin. The most common sources of contamination are cloning vectors (e.g.: plasmid, phage, cosmid, BAC, PAC, YAC) which, unless they are identified and removed, will result in a contaminated sequence. Unintended events can also introduce contamination from other sources such as transposable elements and insertion sequences.

In our case, we observed 16 putative unique proteins (Table 21) which we suspected are results of contamination. Almost all of these 16 proteins were from *P. yoelii* 17XNL (primary host are rodents) where most of their tBLASTN hits were from *Mus musculus* (house mouse) or *Rattus norvegicus* (Norway rat). One *P. inui* Unique is suspected for bacterial contamination. Thus, our primary approach is to prove that these proteins belong to *P. yoelii* 17XNL/*P. inui* and are not contaminat sequences.

For each of these proteins, using PlasmoDBs' BLASTP/tBLASTN tools we performed sequence comparisons as to confirmed/contradict their 'unique' status. Then, using their nucleotide sequence we checked if the TBLASTN hits are possible contaminated sequences using the VecScreen (use of NCBI BLASTN against the UniVec database) and NCBIs' web-BLASTN tools against the reference genomes database (Table 21). We also performed a cross-reference of the mouse tBLASTN hits to verify that are not contaminated sequences.

Our results indicate that 8 out of these 15 "possible contaminated" sequences are, indeed, products of contamination. The VecScreen analysis, using the mouse hits as query, found strong matches to multiple vectors for all these 8 proteins. At the same time, the search of these proteins in PlasmoDB resulted only in self-hits (both for protein and nucleotide sequences) suggesting that these are segments of foreign origin.

Furthermore, for 5 of the remaining proteins we could not find any evidence of contamination and thus we concluded that the mouse/rat hits are False Positive hits. However, their GenBank records were removed and set as Obsolete due to standard genome processing. For only one protein (COGENT ID: PYOE-17XNL-01-6365) we could not find any evidence either that is a product of contamination, obsolete or that it does not belong to P. yoelii 17XNL genome. The PYOE-17XNL-01-6365 is a small peptide of 30 residues long, suggesting that's the reason of being Unique.

**Table 21:** List of the *Plasmodium* unique genes suspected for possible contamination.

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | Status | Notes |
|---|---|---|---|---|---|---|---|
| **1** | PINU-SANT-01-5828 | C922_05875 | 134 | hypothetical protein | N/A | Contamination | Contamination<br>PlasmoDB confirms unique status<br>Blast analysis showed significant similarity only to bacterial genomes |
| **2** | PYOE-17XNL-01-3172 | PY03172 | 127 | hypothetical protein | N/A | Contamination | PlasmoDB confirms unique status<br>VecScreen of mouse hits found strong significant similarity to multiple vectors |
| **3** | PYOE-17XNL-01-6364 | PY06364 | 72 | hypothetical protein | N/A | Obsolete in NCBI | PlasmoDB confirms unique status<br>VecScreen did not found any significant similarity<br>BLASTN against reference genomes confirms mouse hits |
| **4** | PYOE-17XNL-01-6365 | PY06365 | 30 | hypothetical protein | N/A | Strain specific | PlasmoDB confirms unique status<br>tBLASTN hits only mouse-No self-hits<br>VecScreen did not found any significant similarity<br>Missed due to small length |
| **5** | PYOE-17XNL-01-7015 | PY07015 | 99 | hypothetical protein | N/A | Obsolete in NCBI | PlasmoDB tBLASTN found self and *P. gallicaneum* (e-value: $2e^{-28}$)<br>VecScreen found any significant similarity to J02459.1 (1-48502-49; Enterobacteria phage lambda) |
| **6** | PYOE-17XNL-01-7016 | PY07016 | 163 | hypothetical protein | N/A | Contamination | PlasmoDB confirms unique status<br>VecScreen found strong significant similarity to Z22761.1 (1-5585; Retroviral expression vector pSFF DNA) |
| **7** | PYOE-17XNL-01-7503 | PY07503 | 35 | hypothetical protein | N/A | Obsolete in NCBI | Not BLASTP unique (See BLASTP interesting cases)<br>tBLASTN hits in another Plasmodium<br>VecScreen did not found any significant similarity |
| **8** | PYOE-17XNL-01-7580 | PY07580 | 112 | hypothetical protein | N/A | Contamination | PlasmoDB confirms unique status<br>VecScreen did not found any significant similarity |

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | Status | Notes |
|---|---|---|---|---|---|---|---|
| **9** | PYOE-17XNL-01-7581 | PY07581 | 41 | hypothetical protein | N/A | Obsolete in NCBI | BLASTN against reference genomes confirms mouse hits<br>PlasmoDB confirms unique status<br>VecScreen did not found any significant similarity<br>BLASTN against reference genomes confirms mouse hits<br>Missed due to small length |
| **10** | PYOE-17XNL-01-7582 | PY07582 | 60 | hypothetical protein | N/A | Contamination | PlasmoDB confirms unique status<br>VecScreen found strong significant similarity to L07041.1 (673-1123 pMHNeo eukaryotic expression vector) |
| **11** | PYOE-17XNL-01-7604 | PY07604 | 40 | hypothetical protein | N/A | Obsolete in NCBI | BLASTN against reference genomes confirms mouse hits<br>PlasmoDB confirms unique status<br>VecScreen did not found any significant similarity<br>BLASTN against reference genomes confirms mouse hits<br>Missed due to small length |
| **12** | PYOE-17XNL-01-7672 | PY07672 | 84 | hypothetical protein | N/A | Contamination | PlasmoDB confirms unique status<br>VecScreen found strong significant similarity to multiple vectors |
| **13** | PYOE-17XNL-01-7673 | PY07673 | 168 | hypothetical protein | N/A | Contamination | PlasmoDB confirms unique status<br>No tBLASTN self-hits<br>Not BLASTP unique (See BLASTP interesting cases)<br>VecScreen found strong significant similarity to multiple vectors |
| **14** | PYOE-17XNL-01-7674 | PY07674 | 63 | hypothetical protein | N/A | Obsolete in NCBI | PlasmoDB tBLASTN hits in another Plasmodium<br>VecScreen found strong significant similarity to multiple vectors |
| **15** | PYOE-17XNL-01-7676 | PY07676 | 62 | hypothetical protein | N/A | Obsolete | PlasmoDB confirms unique status<br>VecScreen did not found any significant similarity |

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | Status | Notes |
|---|---|---|---|---|---|---|---|
| **16** | PYOE-17XNL-01-7677 | PY07677 | 67 | hypothetical protein | N/A | Contamination | BLASTN against reference genomes confirms mouse hits<br>Missed due to small length<br>PlasmoDB tBLASTN find hit with Pg_2265551.c000322376.Contig1 (Plasmodium_gallinaceum; 6e-09)<br>VecScreen found strong significant similarity to multiple vectors |

**Multiple BLASTP hits**

In this category, we examine those proteins that after the sequence comparison step against the NR database showed significant sequence similarity to more than one sequence (i.e. significant sequence similarity was found to other protein sequences apart from itself). In total, we observed 126 proteins with multiple hits and 15 of these proteins have hits outside the genus of Plasmodium (**Table 22**).

A post processing approach was followed as to establish which hits are True Positives (TP) and show distant homology to the *Plasmodium* proteins or if the significant sequence similarity was due to repeats or CBRs. Thus, for each of these 15 proteins we, first, extracted the FASTA sequences and created a fake BLAST database (i.e. the fake database was composed only by the BLASTP hit sequences). We re-run BLASTP locally using each Unique protein of this category as query against this fake database. Afterwards, we run the stand-alone version of mview (as described in Methods) as to extract the pairwise alignments and create an MSA based on the best local alignments.

In addition, we performed a cross-reference BLASTP using the hits of each protein as query against NR database. This step will ensure which hits are TP and which are spurious due to certain amino acid repeats or CBRs. If the *Plasmodium* proteins are one of the top hits then we consider the hit as TP, otherwise we discard it as FP.

**Table 22:** Summary list of the proteins where significant sequence similarity was found with species outside the genus Plasmodium.

| A/A | COGENT Id | PlasmoDB Id | Protein Length | Annotation | CBRs | #BLASTP hits | #tBLASTN hits | Notes | Status |
|---|---|---|---|---|---|---|---|---|---|
| **1.** | PFRA-NILG-01-4203 | AK88_04244 | 2796 | hypothetical protein | H, T | 2 | 47 | PlasmoDB (BLASTP) missed due to strict e-value (tBLASTN) confirms Plasmodium hits X. laevis cross-reference BLASTP find other Plasmodium as top hits TBLASTN hits with other species (not Plasmodium) discarder due to repeats | Partial |
| **2.** | PINU-SANT-01-5828 | C922_05875 | 134 | hypothetical protein | N/A | 10 | 4 | PlasmoDB confirms unique status Cross-reference of the BLASTP hit confirm homology | Bacterial Contamination |
| **3.** | POVA-CURT-01-1024 | PocGH01_06023700 | 440 | hypothetical protein | T | 37 | 388 | Not a member of OG5_132281 | Partial |

| A/A | COGENT Id | PlasmoDB Id | Protein Length | Annotation | CBRs | #BLASTP hits | #tBLASTN hits | Notes | Status |
|---|---|---|---|---|---|---|---|---|---|
| **4.** | PYOE-17XNL-01-1326 | PY01326 | 81 | hypothetical protein | K | 3 | 13 | Antisense to exons of PY01325 | FP/Obsolete in NCBI |
| **5.** | PYOE-17XNL-01-2162 | PY02162 | 149 | hypothetical protein | N/A | 2 | 18 | PlasmoDB confirms unique status Cross-reference of the BLASTP hit with N. vectensis discarded | Obsolete in NCBI |
| **6.** | PYOE-17XNL-01-4653 | PY04653 | 124 | hypothetical protein | N/A | 32 | 500 | Homologous to 28S rRNA BLASTN using PlasmoDB transcripts | Incorrect gene prediction/Obsolete in NCBI |
| **7.** | PYOE-17XNL-01-5160 | PY05160 | 262 | Unknown, putative | N/A | 5 | 500 | Contains a tandem repeat of 32 aa | Obsolete in NCBI |
| **8.** | PYOE-17XNL-01-5955 | PY05955 | 127 | hypothetical protein | N/A | 8 | 33 | PlasmoDB (tBLASTN) find hits in other Plasmodium species Cross-reference of BLASTP hits only A. limnaeus | Obsolete in NCBI |

| A/A | COGENT Id | PlasmoDB Id | Protein Length | Annotation | CBRs | #BLASTP hits | #tBLASTN hits | Notes | Status |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | (XP_013886717.1) and S. kowalevskii (XP_006818040.1) confirmed TBLASTN mview show good alignments | |
| **9.** | PYOE-17XNL-01-6364 | PY06364 | 72 | hypothetical protein | N/A | 3 | 500 | See **Table 21** | Obsolete in NCBI |
| **10.** | PYOE-17XNL-01-6477 | PY06477 | 58 | hypothetical protein | I | 3 | 500 | PlasmoDB confirms Plasmodium tBLASTN hits C-terminal overlap with 18S rRNA | FP/ Obsolete in NCBI Obsolete in NCBI |
| **11.** | PYOE-17XNL-01-7503 | PY07503 | 35 | hypothetical protein | N/A | 10 | 245 | TBLASTN hits to other Plasmodium species Cross-reference of BLASTP hits confirmed VecScreen did not found any significant | Obsolete in NCBI |

| A/A | COGENT Id | PlasmoDB Id | Protein Length | Annotation | CBRs | #BLASTP hits | #tBLASTN hits | Notes | Status |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | similarity (tBLASTN hits) | |
| **12.** | PYOE-17XNL-01-7604 | PY07604 | 40 | hypothetical protein | N/A | 2 | 500 | See **Table 21** | Obsolete in NCBI |
| **13.** | PYOE-17XNL-01-7673 | PY07673 | 168 | hypothetical protein | N/A | 23 | 500 | See **Table 21** | Contamination |
| **14.** | PYOE-17XNL-01-7674 | PY07674 | 63 | hypothetical protein | N/A | 39 | 500 | PlasmoDB (tBLASTN) find hits to other Plasmodium species Cross-reference of the BLASTP hits confirmed | Obsolete in NCBI |
| **15.** | PYOE-17XNL-01-7676 | PY07676 | 62 | hypothetical protein | N/A | 5 | 500 | See **Table 21** | Obsolete in NCBI |

The analysis of these 15 proteins provided interesting results where 5 protein sequences were excluded due to partial mRNA sequence (obsolete in NCBI), 5 are suspected as products of contamination, 2 are ribosomal encoding genes, 1 as FP Unique due to be the antisense of a missed exon of its neighboring gene and for the remaining proteins, the BLASTP hits from species that do not belong in the genus *Plasmodium* were discarded as FP due to tandem repeats or CBRs.

The PYOE-17XNL-01-4653 (PlasmoDB id: PY04653) is incorrectly predicted as protein coding gene as our analysis demonstrated that is homologous to the 28S rRNA of all completely sequenced *Plasmodium* species currently deposited in PlasmoDB. Specifically, using the genomic sequence provided in PlasmoDBs gene page, we performed a BLASTN search (e-value cutoff: $1e^{-06}$) against the PlasmoDB transcripts database where we observed statistically significant hits to genes annotated as "28S ribosomal RNA" from all *Plasmodium* species. Currently, *P. yoelii* 17XNL has not annotated 28S ribosomal RNA as opposed to the other two *P. yoelii* species that have 4 28S rRNA encoding genes (**Figure 35**).

(A)

Query= MySeq1

Length=375

| | Score | E |
|---|---|---|
| PY04653-t26_1 \| gene=PY04653 \| organism=Plasmodium_yoelii_yoeli... | 677 | 0.0 |
| PY17X_0701151.1 \| gene=PY17X_0701151 \| organism=Plasmodium_yoel... | 677 | 0.0 |
| PY17X_1249360.1 \| gene=PY17X_1249360 \| organism=Plasmodium_yoel... | 677 | 0.0 |
| PYYM_0700900.1 \| gene=PYYM_0700900 \| organism=Plasmodium_yoelii... | 677 | 0.0 |
| PYYM_1248700.1 \| gene=PYYM_1248700 \| organism=Plasmodium_yoelii... | 677 | 0.0 |
| PBANKA_0700761.1 \| gene=PBANKA_0700761 \| organism=Plasmodium_be... | 652 | 0.0 |
| PBANKA_1245861.1 \| gene=PBANKA_1245861 \| organism=Plasmodium_be... | 652 | 0.0 |
| PBANKA_0622961.1 \| gene=PBANKA_0622961 \| organism=Plasmodium_be... | 643 | 0.0 |
| PBANKA_0521261.1 \| gene=PBANKA_0521261 \| organism=Plasmodium_be... | 630 | 1e-179 |
| YYE_04401-t30_1 \| gene=YYE_04401 \| organism=Plasmodium_vinckei_... | 614 | 1e-174 |
| YYG_04280-t30_1 \| gene=YYG_04280 \| organism=Plasmodium_vinckei_... | 614 | 1e-174 |
| PCHAS_1246500.1 \| gene=PCHAS_1246500 \| organism=Plasmodium_chab... | 610 | 1e-173 |
| PY17X_0522600.1 \| gene=PY17X_0522600 \| organism=Plasmodium_yoel... | 609 | 4e-173 |
| PYYM_0521780.1 \| gene=PYYM_0521780 \| organism=Plasmodium_yoelii... | 609 | 4e-173 |
| PCHAS_0937220.1 \| gene=PCHAS_0937220 \| organism=Plasmodium_chab... | 605 | 5e-172 |
| PY17X_0625850.1 \| gene=PY17X_0625850 \| organism=Plasmodium_yoel... | 576 | 2e-163 |
| PYYM_0624920.1 \| gene=PYYM_0624920 \| organism=Plasmodium_yoelii... | 576 | 2e-163 |
| YYE_03725-t30_1 \| gene=YYE_03725 \| organism=Plasmodium_vinckei_... | 572 | 3e-162 |
| YYG_05092-t30_1 \| gene=YYG_05092 \| organism=Plasmodium_vinckei_... | 569 | 4e-161 |
| PCHAS_0624660.1 \| gene=PCHAS_0624660 \| organism=Plasmodium_chab... | 560 | 2e-158 |
| PRELSG_0020800.1 \| gene=PRELSG_0020800 \| organism=Plasmodium_re... | 513 | 2e-144 |
| PCHAS_0521461.1 \| gene=PCHAS_0521461 \| organism=Plasmodium_chab... | 497 | 2e-139 |

(B)

> PY17X_0701151.1 | gene=PY17X_0701151 | organism=Plasmodium_yoelii_yoelii_17X | gene_product=28S ribosomal RNA | transcript_product=28S ribosomal RNA | location=Py17X_07_v3:-(+) | length=3726 | sequence_SO=chromosome | SO=rRNA_encoding | is_pseudo=false

Length=3726

Score = 677 bits (750), Expect = 0.0
Identities = 375/375 (100%), Gaps = 0/375 (0%)
Strand=Plus/Plus

```
Query  1     ATGACAGTAGGACGGTGGTCATGGAAGTCGAAATCCGCTAAGGAGTGTGTAACAACTCAC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1279  ATGACAGTAGGACGGTGGTCATGGAAGTCGAAATCCGCTAAGGAGTGTGTAACAACTCAC  1338

Query  61    CTACCGAATGAACTAGCCCTGAAAATGGATGGCGCTAAAGCGGATTACCGATACCAAGCC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1339  CTACCGAATGAACTAGCCCTGAAAATGGATGGCGCTAAAGCGGATTACCGATACCAAGCC  1398

Query  121   ATAGAAAGAGATTTAAGGGTTTTTTGGGTTCCCCCGACTAATCTTTTTTTGAATCTTTCT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1399  ATAGAAAGAGATTTAAGGGTTTTTTGGGTTCCCCCGACTAATCTTTTTTTGAATCTTTCT  1458

Query  181   ATGAGTAGAAAATCGTGGGGTTTGTGTAGAAGCAAGATGCGTGAGCTCTTGTGGAACATA  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1459  ATGAGTAGAAAATCGTGGGGTTTGTGTAGAAGCAAGATGCGTGAGCTCTTGTGGAACATA  1518

Query  241   TCCCTAGTGCAGATCTTGGTGGAAGTAGCAATTATTCAAATGAGAACTTTGAAGACTGAA  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1519  TCCCTAGTGCAGATCTTGGTGGAAGTAGCAATTATTCAAATGAGAACTTTGAAGACTGAA  1578

Query  301   GTGGAGAAGGGTTTCTTGTCAACTGTGATTGTACAAGAGTTAGCCGCTCCTAAGGGAAAG  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1579  GTGGAGAAGGGTTTCTTGTCAACTGTGATTGTACAAGAGTTAGCCGCTCCTAAGGGAAAG  1638

Query  361   CTGAAATGCGTTTAA  375
             |||||||||||||||
Sbjct  1639  CTGAAATGCGTTTAA  1653
```

**Figure 35: (A)** The results of BLASTN search for the PYOE-17XNL-01-4653 genomic sequence against PlasmoDB transcripts database. **(B)** An example of the matched hit annotation demonstrating that PYOE-17XNL-01-4653 sequence is an 28S rRNA sequence.

One of the most puzzling proteins to analyze in this dataset, was the POVA-CURT-01-1024 (PlasmoDB id: PocGH01_06023700) protein from *P. ovale* curtisi. In PlasmoDB, this protein belongs in an orthologous group (OG5_132281) with proteins annotated as "serine/threonine protein kinase, FIKK family" but we could not find any evidence that this protein should be included in this orthologous group. Instead, our analysis indicated that *P. ovale* has a protein (genomic location is in chromosome 1) which is orthologous to the *P.*

*falciparum* FIKK protein family and that this protein is probably a partial sequence of unknown function. First, using its protein sequence from PlasmoDB gene page (i.e. unmasked protein sequence), we performed a BLASTP search (e-value: $1e^{-6}$) using the integrated BLAST tools of PlasmoDB. The top hits were itself and two proteins from *P. falciparum 3D7* annotated as "serine/threonine protein kinase, FIKK family" (PlasmoDB id: PF3D7_0424700) and "conserved Plasmodium protein, unknown function" (PlasmoDB id: PF3D7_1012800). Consequently, using the PF3D7_0424700 protein sequence we performed a tBLASTN search (e-value: $1e^{-6}$) against *P. ovale* transcripts where the top hit was a protein annotated as serine/threonine protein kinase, FIKK family, putative" (PlasmoDB id: PocGH01_01022500; chromosome 1) sharing 47% sequence identity with the *P. falciparum* protein. Remarkably, we did not observe any statistically significant sequence similarity with the initial *P. ovale* protein suggesting that is not a member of the FIKK protein family. However, when we searched PlasmoDB Annotated Proteins database (e-value cutoff: $1e^{-6}$) using the masked sequence this time, we observed that the top hit in *P. falciparum* 3D7 was a protein annotated as "conserved Plasmodium protein, unknown function" sharing 18% sequence identity. The statistically significant hits in other *Plasmodium* species were partial matches to this *P. ovale* protein as all protein sequences were significantly larger that this protein. These results clearly demonstrate that this protein sequence does not belong in the FIKK protein family and that is probably a partial sequence. Furthermore, all BLASTP hits from species that belong in other genera were discarded as false positives due to CBRs.

**Multiple TBLASTN hits from other genera**

As in the previous category, we observed 32 pan-genome putative unique proteins with tBLASTN hits both in Plasmodium species but also in species from other genera such as human, mouse, zebrafish (**Table 23**). Seven of these proteins had also BLASTP hits in species outside of the *Plasmodium* species, which we analyzed in the previous section. Five proteins were classified as "putative de novo" as our analysis could not find any statistically significant evidence that there are homologues in other species. Furthermore, the tBLASTN hits were discarded as false positives as the significant sequence similarity was in non-coding regions or due to amino acid repeats. The remaining proteins were excluded in the post-processing step due to partial mRNA sequence or because were labeled as obsolete in NCBI NR/NT database and thus, no further analysis were made.

**Table 23:** A list of putative unique proteins with tBLASTN hits outside the genus *Plasmodium*.

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | #BLASTP hits | #TBLASTN hits | Status | Notes |
|---|---|---|---|---|---|---|---|---|---|
| **1.** | PCYN-B-01-2071 | PCYB_092240 | 78 | cyclophilin | N/A | 1 | 43 | Fragment | See **Table** 17 |
| **2.** | PCYN-B-01-5087 | PCYB_002030 | 395 | hypothetical protein | S | 1 | 5 | Partial | Hits from B. bigemina discarded due to CBRs |
| **3.** | PFRA-NILG-01-4203 | AK88_04244 | 2796 | hypothetical protein | H, T | 2 | 141 | Partial | See **Table 22** |
| **4.** | PFRA-NILG-01-594 | AK88_00597 | 109 | hypothetical protein | N/A | 1 | 39 | Partial | Excluded due to partial mRNA |
| **5.** | PINU-SANT-01-2862 | C922_02881 | 114 | hypothetical protein | N | 1 | 161 | Partial | Excluded due to partial mRNA |
| **6.** | PINU-SANT-01-3966 | C922_03996 | 85 | hypothetical protein | N | 1 | 3 | Partial | Excluded due to partial mRNA |
| **7.** | PINU-SANT-01-5828 | C922_05875 | 134 | hypothetical protein | N/A | 1 | 4 | Partial | Excluded due to partial mRNA |
| **8.** | PINU-SANT-01-824 | C922_00829 | 85 | hypothetical protein | N/A | 1 | 68 | Partial | Excluded due to partial mRNA |
| **9.** | PKNO-H-01-2420 | PKNH_1000300 | 178 | hypothetical protein, conserved in P. knowlesi | N/A | 1 | 10 | Partial | Excluded due to partial mRNA |
| **10.** | PKNO-H-01-3988 | PKNH_1304600 | 308 | hypothetical protein | E, K | 1 | 12 | Partial | Excluded due to partial mRNA |
| **11.** | PMAL-UG01-01-254 | PmUG01_02011200 | 229 | hypothetical protein | N | 2 | 42 | Unique | Hits from other genera discarded as FP due to certain amino acid repeats or tBLASTN hits were in intronic regions |

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | #BLASTP hits | #TBLASTN hits | Status | Notes |
|---|---|---|---|---|---|---|---|---|---|
| **12.** | PMAL-UG01-01-4917 | PmUG01_13052300 | 154 | hypothetical protein | S | 3 | 4 | Unique | Hits from other genera discarded as FP due to certain amino acid repeats or tBLASTN hits were in intronic regions |
| **13.** | PMAL-UG01-01-830 | PmUG01_04027400 | 908 | conserved Plasmodium protein, unknown function | E, K, K, R | 3 | 18 | Unique | Hits from other genera discarded as FP due to certain amino acid repeats or tBLASTN hits were in intronic regions |
| **14.** | POVA-CURT-01-1024 | PocGH01_06023700 | 440 | hypothetical protein | T | 37 | 388 | Partial | See **Table 22** |
| **15.** | POVA-CURT-01-6231 | PocGH01_00184800 | 373 | Plasmodium exported protein, unknown function | K | 4 | 1 | Unique | Hits from other genera discarded as FP due to certain amino acid repeats or tBLASTN hits were in intronic regions |
| **16.** | POVA-CURT-01-6717 | PocGH01_00049900 | 1013 | Plasmodium exported protein, unknown function | E, K, K, K | 6 | 1 | Unique | POVA-CURT-01-6231 |
| **17.** | POVA-CURT-01-6890 | PocGH01_00239600 | 483 | conserved Plasmodium protein, unknown function | E, G | 4 | 15 | TRG | See **Table 20** |
| **18.** | PREI-CDC-01-5604 | PRCDC_0050600 | 2133 | hypothetical protein | E, V | 1 | 4 | Partial | Excluded due to partial mRNA |
| **20.** | PVIN-PECR-01-2106 | YYG_02134 | 125 | hypothetical protein | K | 3 | 4 | TRG | **Table 20** |
| **21.** | PVIN- | YYG_02959 | 188 | hypothetical | A | 1 | 1 | Partial | Excluded due to partial |

| A/A | COGENT ID | PlasmoDB ID | Length | Annotation | CBRs | #BLASTP hits | #TBLASTN hits | Status | Notes |
|---|---|---|---|---|---|---|---|---|---|
| | PECR-01-2925 | | | protein | | | | | mRNA |
| **22.** | PVIV-Sal1-01-3012 | PVX_113960 | 116 | hypothetical protein | N/A | 1 | 15 | Partial | Excluded due to partial mRNA |
| **23.** | PVIV-Sal1-01-3908 | PVX_118075 | 122 | hypothetical protein | K | 1 | 8 | Partial | Excluded due to partial mRNA |
| **24.** | PYOE-17XNL-01-1326 | PY01326 | 81 | hypothetical protein | K | 3 | 13 | FP/Obsolete in NCBI | See **Table 22** |
| **25.** | PYOE-17XNL-01-1687 | PY01687 | 171 | hypothetical protein | N | 1 | 32 | Partial | Excluded due to partial mRNA |
| **26.** | PYOE-17XNL-01-1930 | PY01930 | 76 | hypothetical protein | N/A | 1 | 120 | Partial | Excluded due to partial mRNA |
| **27.** | PYOE-17XNL-01-3022 | PY03022 | 76 | hypothetical protein | N/A | 1 | 11 | Partial | Excluded due to partial mRNA |
| **28.** | PYOE-17XNL-01-3789 | PY03789 | 73 | hypothetical protein | F | 1 | 15 | Obsolete in NCBI | Excluded due to partial mRNA |
| **29.** | PYOE-17XNL-01-4653 | PY04653 | 124 | hypothetical protein | N/A | 32 | 500 | Incorrect gene prediction/Obsolete in NCBI | See **Table 22** |
| **30.** | PYOE-17XNL-01-5160 | PY05160 | 262 | Unknown, putative | N/A | 5 | 500 | Obsolete in NCBI | See **Table 22** |
| **31.** | PYOE-17XNL-01-5955 | PY05955 | 127 | hypothetical protein | N/A | 8 | 33 | Obsolete in NCBI | See **Table 22** |
| **32.** | PYOE-17XNL-01-6477 | PY06477 | 58 | hypothetical protein | I | 3 | 500 | FP/ Obsolete in NCBI Obsolete in NCBI | See **Table 22** |

**Concluding remarks**

In this section we described *in-depth comparative genomics* of malaria parasites putative unique proteins, also examining the roles of CBRs. We started our analysis with 1201 unique protein candidates where, through exhaustive "filtering", we eliminated partial/fragmented protein sequences, detected gene prediction/functional annotation artifacts and identified contaminant sequences. Our carefully conducted bioinformatics analysis demonstrated that only a small subset of the initial putative unique proteins (96) does not have any detectable homologs in *Plasmodium* pan-genome or in other lineages and can be considered as genuinely *de novo* genes.

Among the initial putative unique and *Plasmodium* species/strain specific proteins, we identified 25 TRGs illuminating the importance of our semi-manually conducted analysis. We must note, however, that these 25 TRGs are only a subset of the true complement of *Plasmodium* TRGs which definitely deserve a more detailed analysis.

Furthermore, among the initial putative *de novo* proteins we identified an unusually short protein with significant similarity to a subunit of the OST complex, which lead us to challenge the established view for the phylogenomic distribution of this key eukaryotic complex, as described in the next section.

### 3.4. OST complex subunits in protists

**Identification of putative genes encoding Ost4p across Plasmodium spp.**

The Ost4p subunit was first described in vanadate-resistant yeast mutants, which were defective in early steps of N-linked glycosylation and was characterized as an unusually small subunit of the OST complex (Chi et al., 1996). It was later shown that the yeast Ost4p is important in the incorporation of Ost3p/Ost6p into the complex (Spirig et al., 2005) and, more recently, its importance in the assembly of the human OST complex was also demonstrated (Dumax-Vorzet et al., 2013). The recently reported, atomic resolution, three dimensional (3D) structures of the yeast (Bai et al., 2018; Wild et al., 2018) and the canine (Braunger et al., 2018) OST complex point to a possible scaffolding function for this subunit.

Systematic literature search (see Data and Methods) did not reveal any works mentioning any evidence for Ost4p being present in any *Plasmodium* species. When querying the PlasmoDB online resource (last accessed: June 26, 2018) with a "Gene Text Search" using the keyword 'ost4' three entries were retrieved:

a. YYG_02734 (*P. vinckei petteri* strain CR) and YYE_03331 (*P. vinckei petteri* strain vinckei): correspond to intron-less transcripts encoding two proteins predicted to consist of 517 and 518 amino acid residues respectively. These protein sequences show significant similarity to sequences of Wbp1p/OST48 subunits, as evidenced by their matches to the specific profile Hidden Markov Model (pHMM) in the PFAM database (PFAM: PF03345; e-value: 2.8E-53 for both sequences). Thus, their annotation in PlasmoDB as 'oligosaccharyl transferase complex subunit OST4' is obviously incorrect. Notably, both the above mentioned PlasmoDB entries are predicted to belong in the same orthologous group (OrthoMCL ID: OG5_128922) and their syntenic homologs in other *Plasmodium* species are (correctly) annotated as putative dolichyl-diphosphooligosaccharide–protein glycosyltransferase 48 kDa subunits.

b. AK88_01616 (*P. fragile* strain nilgiri): corresponds to a 108bp transcript (residing in supercontig KQ001658), which encodes for two exons with a predicted 35 amino acid residue hypothetical protein product (Supplementary text). This gene is found in a region highly conserved in all *Plasmodium* strains/species present in PlasmoDB (Figure 36), however no homologs for this protein are reported in other *Plasmodium* species. Importantly, this protein sequence shows significant similarity to the PFAM pHMM (PFAM: PF10215; e-value: 4.9E-05) for Ost4.

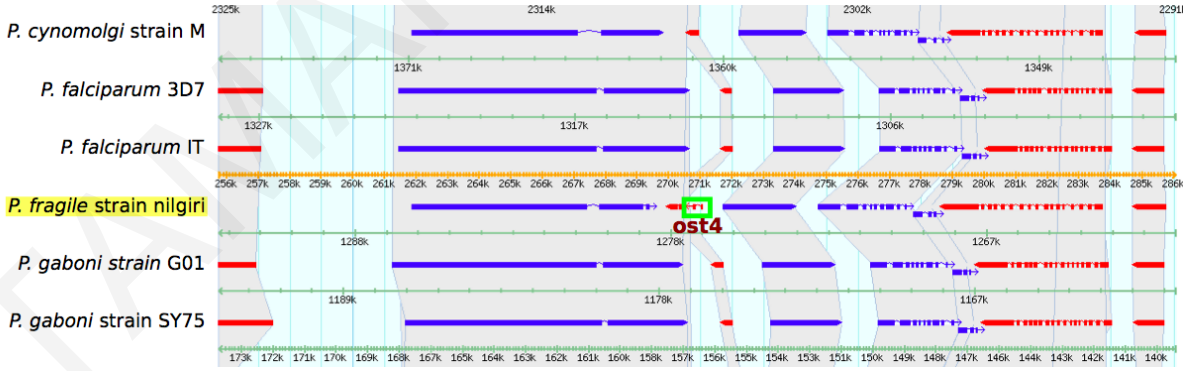



**Figure 36:** Syntenic region surrounding gene AK88_01616, encoding a putative OST4 subunit in P. fragile strain nilgiri. The green box indicates the putative ost4 gene lying within a region of conserved orthologs among all Plasmodium genomes. An enlarged figure of the syntenic region depicting all *Plasmodium spp.* is in Appendix II (**Figure 52**).

*Sequence-based database search and ost4 gene prediction:* Subsequently, using as query the *P. fragile* ost4 gene and protein sequences, we set to identify genes encoding ost4 homologs across other plasmodial species with completely sequenced genomes. When performing a BLASTP search (protein sequence query against the database of all protein sequences

known/predicted to be encoded in *Plasmodium* species) we were unable to identify any significant sequence similarity –apart from the trivial hit to the query sequence itself, even with very permissive parameter settings (e-value cutoff: 10.0; no filter for low complexity sequences). However, a tBLASTN search against six-frame translations of plasmodial genomic sequences returned several hits among all the plasmodial species/strains examined (Table 24).

**Table 24:** Identification of putative ost4 orthologs in *Plasmodium* species. Detailed results of a tBLASTN search against all genomic data in PlasmoDB, using as query the sequence AK88_01616 from *P. fragile* strain nilgiri. Significant e-values in bold typeset. It is worth mentioning that in most species/strains the hits clearly indicate a similar gene structure with the CDS split among two exons. The genomic coordinates of all hits correspond to the intergenic region within the syntenic block in which AK88_01616 resides. The hits in *P. fragile* strain nilgiri correspond to the currently annotated CDS, with a single residue overlap.

| OST4-tBLASTn-chromosomes | | tBLASTn hits | | | | | | | | Query | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Strain | Syntenic Chromosome/Contig | Significant (E<0.001) | E-value | Bit score | %ID | %POS | %GAPS | From | To | From | To |
| Plasmodium berghei ANKA | PbANKA_14_v3 | No | 3.00E+01 | 25.0 | 77 | 100 | 0 | 1830688 | 1830726 | 1 | 13 |
| Plasmodium chabaudi chabaudi | PCHAS_14_v3 | No | 3.00E+01 | 25.0 | 77 | 100 | 0 | 1827105 | 1827143 | 1 | 13 |
| | | No | 3.90E+01 | 24.6 | 48 | 72 | 7 | 1827282 | 1827362 | 5 | 33 |
| Plasmodium coatneyi Hackeri | CP016252 | Yes | **6.00E-05** | 42.7 | 83 | 91 | 0 | 1016312 | 1016244 | 13 | 35 |
| | | No | 1.00E+00 | 29.6 | 92 | 100 | 0 | 1016542 | 1016504 | 1 | 13 |
| Plasmodium cynomolgi strain B | DF157106 | No | 2.00E-03 | 38.1 | 74 | 83 | 0 | 2212791 | 2212859 | 13 | 35 |
| | | No | 5.80E+00 | 27.3 | 92 | 100 | 0 | 2212526 | 2212564 | 1 | 13 |
| Plasmodium falciparum 3D7 | Pf3D7_12_v3 | No | 8.30E-01 | 30.0 | 82 | 94 | 0 | 1358670 | 1358720 | 13 | 29 |
| | | No | 1.10E+01 | 26.6 | 73 | 93 | 0 | 1358501 | 1358545 | 1 | 15 |
| Plasmodium falciparum IT | PfIT_12_v3 | No | 8.30E-01 | 30.0 | 82 | 94 | 0 | 1309909 | 1309959 | 13 | 29 |
| | | No | 1.10E+01 | 26.6 | 73 | 93 | 0 | 1309740 | 1309784 | 1 | 15 |
| Plasmodium fragile strain nilgiri | KQ001658 | Yes | **2.00E-06** | 47.4 | 100 | 100 | 0 | 270831 | 270763 | 13 | 35 |
| | | No | **3.80E-01** | 31.2 | 100 | 100 | 0 | 271065 | 271027 | 1 | 13 |
| Plasmodium gaboni strain SY75 | CM003867.1 | No | 1.90E-02 | 30.4 | 82 | 94 | 0 | 1174423 | 1174473 | 13 | 29 |
| | | No | 1.90E-02 | 26.2 | 77 | 100 | 0 | 1174280 | 1174318 | 1 | 13 |
| Plasmodium gallinaceum 8A | PGAL8A_union_v1_archived_contig_130 | No | 2.70E+01 | 25.4 | 69 | 100 | 0 | 153373 | 153411 | 1 | 13 |
| Plasmodium inui San Antonio 1 | KI965474 | Yes | **2.00E-04** | 41.2 | 78 | 87 | 0 | 214882 | 214950 | 13 | 35 |
| | | No | 6.20E+00 | 27.3 | 92 | 100 | 0 | 214648 | 214686 | 1 | 13 |
| Plasmodium knowlesi strain H | PKNH_14_v2 | Yes | **5.00E-05** | 42.7 | 87 | 96 | 0 | 2292317 | 2292385 | 13 | 35 |
| | | No | 1.00E+00 | 29.6 | 92 | 100 | 0 | 2292068 | 2292106 | 1 | 13 |
| Plasmodium malariae UG01 | PmUG01_14_v1 | No | 1.30E+01 | 26.6 | 85 | 92 | 0 | 2676380 | 2676418 | 1 | 13 |
| Plasmodium ovale curtisi GH01 | PocGH01_14_v1 | No | 6.40E-01 | 27.7 | 75 | 94 | 0 | 2083015 | 2083062 | 1 | 16 |
| | | No | 6.40E-01 | 22.7 | 56 | 78 | 0 | 2083178 | 2083231 | 13 | 30 |
| Plasmodium reichenowi CDC | PrCDC_12_v3 | No | 1.30E-01 | 32.3 | 36 | 42 | 57 | 1300305 | 1300505 | 1 | 29 |
| Plasmodium relictum SGS1-like | PRELSG_14_v1 | No | 1.80E+01 | 25.8 | 77 | 100 | 0 | 1867450 | 1867488 | 1 | 13 |
| Plasmodium vinckei petteri strain CR | KI965394 | No | 3.00E+01 | 25.0 | 77 | 100 | 0 | 1824222 | 1824260 | 1 | 13 |
| Plasmodium vinckei vinckei strain vinckei | KL446945 | No | 2.40E+00 | 25.0 | 77 | 100 | 0 | 626485 | 626447 | 1 | 13 |
| | | No | 2.40E+00 | 23.5 | 48 | 69 | 7 | 626337 | 626257 | 5 | 33 |
| Plasmodium vivax P01 | PvP01_14_v1 | Yes | **5.00E-04** | 39.7 | 76 | 86 | 0 | 2225148 | 2225210 | 13 | 33 |
| | | No | 1.00E+00 | 29.6 | 92 | 100 | 0 | 2224884 | 2224922 | 1 | 13 |
| Plasmodium vivax Sal-1 | Pv_Sal1_chr14 | Yes | **5.00E-04** | 39.7 | 76 | 86 | 0 | 2216470 | 2216532 | 13 | 33 |
| | | No | 1.00E+00 | 29.6 | 92 | 100 | 0 | 2216206 | 2216244 | 1 | 13 |
| Plasmodium yoelii yoelii 17X | Py17X_14_v3 | No | 3.00E+01 | 25.0 | 77 | 100 | 0 | 1968466 | 1968504 | 1 | 13 |
| Plasmodium yoelii yoelii 17XNL | AABL01001442 | No | 3.00E+01 | 25.0 | 77 | 100 | 0 | 15057 | 15095 | 1 | 13 |
| Plasmodium yoelii yoelii YM | PyYM_14_v1 | No | 3.00E+01 | 25.0 | 77 | 100 | 0 | 1948015 | 1948053 | 1 | 13 |

For all *Plasmodium* species/strains (including those for which only insignificant hits were found) we used the FGENESH+ web interface for performing gene prediction (Table 25and Supplementary text). Predictions for 21 out of the 22 examined *Plasmodium* species/strains returned two-exon genes, their lengths (from the predicted transcription start site to the predicted polyadenylation signal) ranging from 432 bps (*P. berghei* ANKA) to 1106 bps (*P. inui* san Antonio 1), with a similar structure: the coding region of the first exon was invariably 37 bps long, whereas the coding region on the second exon varied in length from 59 bps to 71 bps (Table 25). Remarkably, for *P. ovale curtisi* GH01 a two-exon gene was

predicted, however no polyadenylation site was detected in the examined genomic segment.

**Table 25:** Prediction of ost4 genes in Plasmodium species. FGENESH+ 2.6 gene predictions using the *P. falciparum* gene model and assisted by the protein sequence of the annotated ost4 from *P. fragile* strain nilgiri (PlasmoDB: AK88_01616). In two cases where the initially predicted gene models seemed less reliable (*P. chabaudi chabaudi* and P. relictum SGS1-like) we report predictions assisted by a template protein sequence from an evolutionary related species (*P. yoelii* YM and *P. falciparum* 3D7 respectively). All predicted genes have their coding sequences split among two exons, with remarkable gene structure similarity. For the gene predicted in P. ovale curtisi GH01 no poly-adenylation site was predicted within the examined genomic region.

| FGENESH+ 2.6 gene predictions (ost4) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Genomic region analysed | | | | Gene Predictions | | | | | | | | |
| Species/Strain | Template | Chromosome/Contig | Start | End | Strand | TSS | CDS1-start | CDS1-end | CDS2-start | CDS2-end | PolyA-site | CDS1-length | CDS2-length | Full-length |
| Plasmodium berghei ANKA | ost4p P. fragile | 14 | 1830207 | 1831206 | + | 1830620 | 1830688 | 1830724 | 1830913 | 1830971 | 1831051 | 37 | 59 | 432 |
| Plasmodium chabaudi chabaudi | ost4p P. yoelii YM | 14 | 1825642 | 1827641 | + | 1826550 | 1827105 | 1827141 | 1827307 | 1827365 | 1827439 | 37 | 59 | 890 |
| Plasmodium coatneyi Hackeri | ost4p P. fragile | CP016252 | 1015393 | 1017392 | - | 1016859 | 1016506 | 1016542 | 1016241 | 1016311 | 1015916 | 37 | 71 | 944 |
| Plasmodium cynomolgi strain B | ost4p P. fragile | DF157106 | 2212193 | 2213692 | + | 2212685 | 2213026 | 2213062 | 2213292 | 2213362 | 2213642 | 37 | 71 | 958 |
| Plasmodium falciparum 3D7 | ost4p P. fragile | 14 | 1358195 | 1359194 | + | 1358290 | 1358501 | 1358538 | 1358671 | 1358732 | 1358861 | 38 | 62 | 572 |
| Plasmodium falciparum IT | ost4p P. fragile | 12 | 1309434 | 1310433 | + | 1309529 | 1309740 | 1309776 | 1309910 | 1309971 | 1310100 | 37 | 62 | 572 |
| Plasmodium gaboni strain SY75 | ost4p P. fragile | CM003867.1 | 1173877 | 1174876 | + | 1174090 | 1174280 | 1174316 | 1174424 | 1174485 | 1174610 | 37 | 62 | 521 |
| Plasmodium gallinaceum 8A | ost4p P. fragile | PGAL8A_union_v1_archived_contig_130 | 152892 | 153891 | + | 153015 | 153373 | 153409 | 153533 | 153600 | 153820 | 37 | 68 | 806 |
| Plasmodium inui San Antonio 1 | ost4p P. fragile | KI965474 | 214299 | 215798 | + | 214350 | 215148 | 215184 | 215383 | 215453 | 215455 | 37 | 71 | 1106 |
| Plasmodium knowlesi strain H | ost4p P. fragile | 14 | 2291727 | 2292726 | + | 2291768 | 2292068 | 2292104 | 2292318 | 2292388 | 2292461 | 37 | 71 | 694 |
| Plasmodium malariae UG01 | ost4p P. fragile | 14 | 2675590 | 2677589 | + | 2675942 | 2676380 | 2676416 | 2676751 | 2676818 | 2676844 | 37 | 68 | 903 |
| Plasmodium ovale curtisi GH01 * | ost4p P. fragile | 14 | 2082623 | 2084122 | + | 2082674 | 2083515 | 2083551 | 2083679 | 2083726 | N/A | 37 | 48 | 1053 |
| Plasmodium reichenowi CDC | ost4p P. fragile | 12 | 1299905 | 1300904 | + | 1300098 | 1300305 | 1300341 | 1300456 | 1300517 | 1300639 | 37 | 62 | 542 |
| Plasmodium relictum SGS1-like | ost4p P. falciparum 3D7 | 14 | 1866969 | 1867968 | + | 1867162 | 1867450 | 1867486 | 1867614 | 1867681 | 1867883 | 37 | 68 | 722 |
| Plasmodium vinckei petteri strain CR | ost4p P. fragile | KI965394 | 1823851 | 1824850 | + | 1823959 | 1824222 | 1824258 | 1824425 | 1824483 | 1824557 | 37 | 59 | 599 |
| Plasmodium vinckei vinckei strain vinckei | ost4p P. fragile | KL446945 | 625871 | 626870 | - | 626716 | 626449 | 626485 | 626254 | 626312 | 625952 | 37 | 59 | 765 |
| Plasmodium vivax P01 | ost4p P. fragile | 14 | 2224547 | 2226046 | + | 2225189 | 2225384 | 2225420 | 2225649 | 2225719 | 2225901 | 37 | 71 | 713 |
| Plasmodium vivax Sal-1 | ost4p P. fragile | 14 | 2215869 | 2217368 | + | 2216645 | 2216706 | 2216742 | 2216971 | 2217041 | 2217223 | 37 | 71 | 579 |
| Plasmodium yoelii yoelii 17X | ost4p P. fragile | 14 | 1967985 | 1968984 | + | 1968165 | 1968466 | 1968502 | 1968719 | 1968777 | 1968905 | 37 | 59 | 741 |
| Plasmodium yoelii yoelii 17XNL | ost4p P. fragile | AABL01001442 | 14576 | 15575 | + | 14756 | 15057 | 15093 | 15310 | 15368 | 15496 | 37 | 59 | 741 |
| Plasmodium yoelii yoelii YM | ost4p P. fragile | 14 | 1947663 | 1948662 | + | 1947714 | 1948015 | 1948051 | 1948268 | 1948326 | 1948454 | 37 | 59 | 741 |
| | | | | | | | | | | | | | | |
| Notes | | | | | | | | | | | | | | |
| * no polyA site detected | | | | | | | | | | | | | | |

The genomic location of all predicted ost4 genes occurred in predicted intergenic regions, surrounded by syntenic orthologues (Syntenic Neighborhood Conservation Index: SNCI = 1), suggesting that the hits possibly correspond to underpredictions by gene-finding pipelines. This is by no means surprising, since it is known that short ORFs or protein coding genes are often overlooked, in order to avoid spurious gene calls (Delcourt et al., 2018; Kessler et al., 2003).

Searching against RNAseq experimental data from *Plasmodium* species deposited in the NCBI Sequence Read Archive (SRA) database with BLASTN, we managed to identify reads (corresponding to expressed mRNAs) which significantly match with the CDS of the *P. fragile* ost4 gene. For example, a BLASTN search against SRA run SRR3274045 (a *P. falciparum* RNAseq experimental run, with 125b long paired-end Illumina reads) returns a highly significant match to sequence 2180432.1 (Score: 68.0 bits; evalue: 2E-10; identities: 77%; gaps: 0%). Performing a BLASTN search using as query the identified read against the *P. falciparum* genome, returns a significant match (100% identity) with two segments corresponding to the coding sequence previously predicted in chromosome 12 (Figure 37a, 37c). Importantly, the paired read (2180432.2) aligns in the region upstream the predicted transcription start site (Figure 37b, 37c), indicating that further work is needed to determine the precise extent and sequence of the actual ost4 transcript.

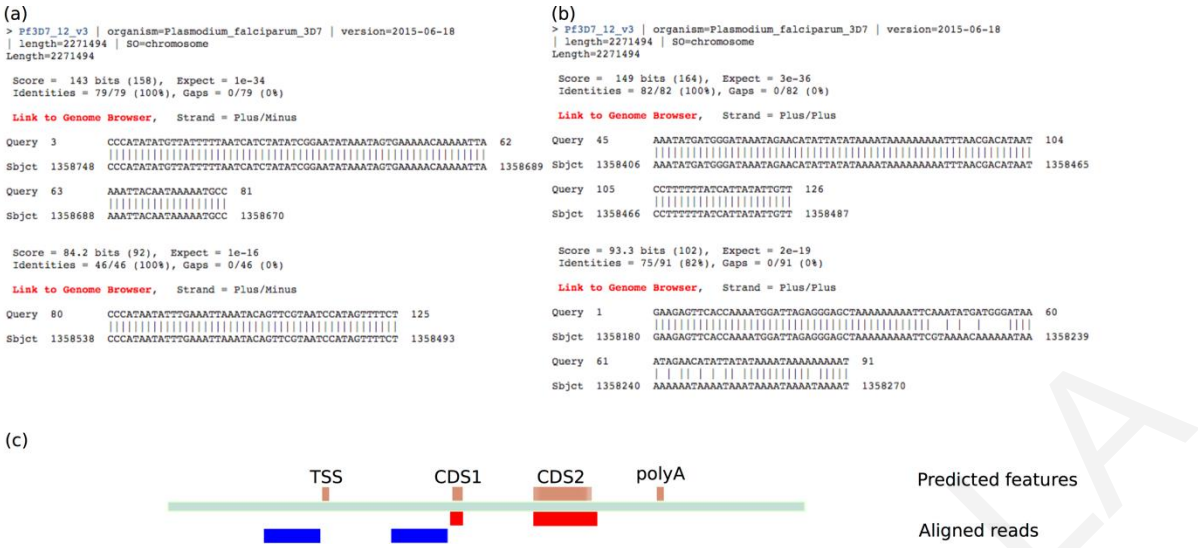



**Figure 37:** Transcriptomic RNA sequencing data supporting the expression of ost4 in *P. falciparum*. (a) Sequence read SRA|SRR3274045.2180432.1 from *P. falciparum* (125bps long) completely matched the genomic region containing the coding exons for the predicted ost4 gene in the *P. falciparum* 3D7 genome. (b) The paired-end read (SRA|SRR3274045.2180432.2) partially overlaps this genomic region which –along with (a) supports that the newly identified gene is indeed transcribed. (c) A graphical depiction of ost4 gene features as predicted by FGENESH+ (brown box) and their support by RNAseq data (red boxes: read SRA|SRR3274045.2180432.1; blue boxes: read SRA|SRR3274045.2180432.2). Cartoon not drawn to scale.

**Identification of putative genes encoding Ost3p/Ost6p across Plasmodium spp.**

Ost3p was among the first identified components of the yeast OST (Kelleher and Gilmore 1994). Ost6p was identified as a possible Ost3p paralog (through low sequence similarity) and characterized as a subunit of the OST complex (Knauer and Lehle, 1999). Single OST3/OST6 deletion mutants lead to mild hypoglycosylation in vivo, whereas double mutants demonstrate more severe hypoglycosylation (Knauer and Lehle, 1999). More recent findings corroborate the view that Ost3p and Ost6p participate in the OST complex in a mutually exclusive manner (Wild et al., 2018). Vertebrate genomes encode two putative homologs of Ost3p/Ost6p (namely Tusc3 and MagT1), while fully sequenced nematodes, fungi (except *Encephalitozoon cuniculi*), arthropods, plants and some protists (such as *Cryptosporidium parvum* but not *Plasmodium spp.*) were reported to encode at least one member of this family (Kelleher and Gilmore, 2006).

Systematic literature search (see Data and Methods) did not identify any works mentioning the existence of any evidence for Ost3p/Ost6p being present in any *Plasmodium* species. When querying the PlasmoDB online resource (last accessed: June 26, 2018) with a “Gene Text Search” using the keywords ‘ost3’ or ‘ost6’ retrieved exactly one gene/protein entry per genome (with the notable exception of *P. yoelii* yoelii 17XNL –see next, where no entry

was found); the protein products of these entries were most often described with generic terms, such as 'conserved Plasmodium membrane protein, unknown function' and in a few cases as 'OST3/OST6 domain-containing protein, putative'. All these protein products are annotated with the PFAM signature OST3_OST6 domain (PFAM: PF04756), their lengths ranging between 376-551 aa residues. These entries are encoded in a highly syntenic genomic region (SNCI=1, for all species/strains except for *P. fragile* strain nilgiri with SNCI=0.9) and share a conserved gene architecture (all genes are single-exon, with the exception of YYG_04526-t30_1 and YYE_04522-t30_1 from *P. vinckei petteri strain CR* and *P. vinckei vinckei strain vinckei* with 2 and 3 exons respectively). Moreover, they are reported to belong in the same orthologous family (OrthoMCL ID: OG5_166992), thus providing strong evidence they are bona fide Ost3p/Ost6p orthologs.

Combined prediction of transmembrane topology and signal peptides using TOPCONS (Tsirigos et al., 2015) yields results similar to the experimentally determined topology for Ost3p/Ost6p. In most of the Plasmodium Ost3p/Ost6p candidate sequences identified here, 5 transmembrane helices are predicted, one near the N-terminus and four in the C-terminal region of the polypeptide. Available structural data indicate that only the four transmembrane helical segments are integral to the ER membrane, while the N-terminal region possesses a cleavable signal peptide, with the mature polypeptide adopting a topology with its termini facing the ER-lumen (Figure 4b). Apparently, the N-terminal signal peptide of the pre-protein is being confused to a transmembrane helix, a common pitfall of transmembrane topology prediction tools (Reynolds et al., 2008), which might also be aggravated by peculiar features of secretory signal peptides in *Plasmodium* (Römisch, 2005; Tonkin et al., 2006) .

In addition, in order to identify a putative ortholog in *P. yoelii* yoelii 17XNL, we performed a tBLASTN sequence against its genome in PlasmoDB using as query the predicted Ost3p/Ost6p from its close relative *P. yoelii* yoelii YM (PlasmoDB ID: PYYM_0208600). A significant hit covering the whole query sequence (E-value: 0; identities: 100%) was detected in contig AABL01002061 (ranging from 8897-10180 in the minus strand). This genomic region overlaps with a predicted gene encoding a 176 aa long protein (PlasmoDB ID: PY06178), with sequence similarity throughout its length to the central region of other plasmodial Ost3p/Ost6p subunits (data not shown), indicating a case of incorrect gene fragment prediction. Transcriptomic data available in PlasmoDB for *P. berghei*, *P.*

*falciparum*, *P. vivax* and *P. yoelii* further support that the respective predicted genes (PlasmoDB IDs: PBANKA_0205700, PF3D7_0107700, PVP01_0207200, PY17X_0207100) are actually expressed.

It is worth mentioning that, during the conclusion of this work, we came across a recently published work reporting on the characterization of Pb51 (corresponding to the PBANKA_0205700 entry) as a putative Ost3p/Ost6p homolog (Wang et al., 2017). According to this work, Pb51 often localizes on outer surface of *P. berghei*, suggesting that these proteins may play additional roles in *Plasmodium.*

**Identification of putative genes encoding Ost5p across Plasmodium spp.**

Ost5p is another small subunit of the OST complex, which was among the first components of the yeast complex to be characterized (Kelleher and Gilmore, 1994) and its deletion was later shown to lead to reduced OST activity, even though it is not essential for growth (Reiss et al., 1997). More recently, its human counterpart (TMEM258) has been characterized (Blomen et al., 2015) and it was demonstrated that its depletion results in reduced N-linked glycosylation (Graham et al., 2016). With the elucidation of the atomic resolution structures of the yeast and canine OST complexes, a corrected transmembrane topology was proposed for Ost5p and TMEM258 and has highlighted its role in the appropriate assembly of the complex (Braunger et al., 2018; Wild et al., 2018).

Our literature search did not reveal any reported evidence for an Ost5p homolog within *Plasmodium*; moreover, a PlasmoDB 'Gene Text Search' using either the 'Ost5' or the 'TMEM258' keywords returned no results. However, performing an advanced UniProt search for entries within the genus 'Plasmodium' with the keyword 'Ost5' anywhere in the text, returned 21 unreviewed (i.e. automatically annotated) entries (Supplement) from different *Plasmodium* strains/species: they are all relatively short polypeptides (83 residues long) and annotated as 'Uncharacterized protein'. All the above entries match to the PFAM pHMM for Ost5 (PFAM: PF05251).

A BLASTP search against protein sequences in PlasmoDB using as query the sequence retrieved from *Plasmodium falciparum* 3D7 (UniProt ACC: C6S3L2) yielded one significant hit per species/strain (e-values < 1E-18), with only a few exceptions (see below). All the detected gene products encode 83 aa residue long polypeptides, residing in a conserved syntenic region (all SNCIs ≥ 0.8), predicted to span the membrane bilayer twice, consistent

with our current knowledge on the transmembrane topology of the Ost5 subunit. Additionally, transcriptomic data available in PlasmoDB for *P. berghei*, *P. falciparum*, *P. vivax* and *P. yoelii* further support that the respective predicted genes (PlasmoDB IDs: PBANKA_1456600, PF3D7_1243200, PVP01_1460300, PY17X_1459100) are actually expressed.

For *P. coatneyi hackeri*, *P. fragile strain* nilgiri, *P. vinckey* (petteri strain CR and vinckey strain vinckey) and *P. yoelii yoelii* 17XNL no protein sequence in the database yielded significant results. We resorted to a tBLASTN search with the putative *P. falciparum* 3D7 Ost5p sequence to identify 1 significant match to a continuous genomic segment in each of these species, suggesting that Ost5p is universally present in all plasmodial genomes (Table 26).

**Table 26:** Prediction of ost5 genes in Plasmodium species. FGENESH+ 2.6 gene predictions using the P. falciparum gene model and assisted by the protein sequence of the P. falciparum 3D7 protein sequence with a match to the PFAM pHMM for Ost5 (UniProt ACC: C6S3L2). Gene prediction was performed only for those strains were no protein entry matched with the P. falciparum 3D7 sequence in a BLASTP search. All the gene models correspond to 83 residue long polypeptides.

| FGENESH+ 2.6 gene predictions (ost5) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Genomic region analysed | | | | Gene Predictions | | | | | |
| Species/Strain | Template | Chromosome/Contig | Start | End | Strand | TSS | CDS-start | CDS-end | PolyA-site | CDS-length | Full-length |
| Plasmodium coatneyi Hackeri | ost5p P. falciparum 3D7 | CP016252 | 515943 | 516942 | - | 516837 | 516316 | 516567 | 516089 | 252 | 749 |
| Plasmodium fragile Nilgiri | ost5p P. falciparum 3D7 | KQ001675 | 123321 | 124320 | - | 124233 | 123694 | 123945 | 123679 | 252 | 555 |
| Plasmodium inui San Antonio 1 | ost5p P. falciparum 3D7 | KI965487 | 196199 | 197198 | - | 197147 | 196569 | 196820 | 196547 | 252 | 601 |
| Plasmodium vinckei petteri strain CR | ost5p P. falciparum 3D7 | KI965394 | 2162096 | 2163095 | + | 2162387 | 2162472 | 2162723 | 2162982 | 252 | 596 |
| Plasmodium vinckei vinckei strain vinckei | ost5p P. falciparum 3D7 | KL446945 | 289519 | 290518 | - | 290242 | 289892 | 290143 | 289629 | 252 | 614 |
| Plasmodium yoelii yoelii 17XNL | ost5p P. falciparum 3D7 | AABL01000332 | 6326 | 7325 | - | 7112 | 6699 | 6950 | 6456 | 252 | 657 |

## Identification of missing OST subunits in other protists

In order to bring our results within an evolutionary context, we expanded our analysis to identify missing oligosaccharyltransferase subunits in select protist species. In particular, we focused on the alveolate *Cryptosporidium parvum* (phylum: Apicomplexa), the excavate *Trichomonas vaginalis* (phylum: Metamonada), and *Entamoeba histolytica* (from the Amoebozoa supergroup) as sample species of this diverse group of unicellular eukaryotes. These species have been reported to encode simple OST complexes composed of 6, 4 and 4 subunits respectively (Kelleher and Gilmore, 2006). All these protists lacked the Ost5 and Swp1 subunits, while *T. vaginalis* and *E. histolytica* additionally lacked Ost3/6 and Ost4, thus displaying similar subunit composition with the previously aggreed composition of the *Plasmodium* OST complex.

Following the same search strategy that we used for the identification of plasmodial OST subunits, we managed to detect all missing subunits from *C. parvum*, putative Ost3/6 subunits in both *T. vaginalis* and *E. histolytica* and a putative Swp1 homolog in *T. vaginalis* (Figure 4, Appendix III-Supplementary text). Therefore, these results collectively indicate that it is necessary to update our view on the subunit composition of OST complexes in protists in general.

**Conservation of OST subunits**

*Ost4*: The predicted amino acid sequences for *Plasmodium* Ost4 share significant pairwise sequence similarity to each other, ranging from approximately 55% to 100% sequence identity. We generated a multiple sequence alignment of the plasmodial Ost4p sequences with their predicted homolog in *Cryptosporidium parvum* (strain Iowa II) and the homologs in yeast (UniProt AC: Q99380) and human (UniProt AC: P0C6T2), for which structural data exist (Figure 38).

Even though most plasmodial Ost4p sequences share less than 30% identity to their aligned homologs from other species (data not shown), their predicted helical TM region is almost perfectly aligned to those deduced from the available 3D structures (Figure 38). The same holds for the sequence from *C. parvum*, with pairwise identities approximately 20% to the yeast and <20% to the human counterparts, while maintaining relatively high sequence identity to its *Plasmodium* homologs (~30-50% identities). Deletions are observed in the short N- and C-terminal regions, with more variability (both in terms of length and amino acid conservation) in the cytosolic C-terminal region (Figure 38).

We also observe a glycine residue residing within the transmembrane helix (GLY15 in the yeast sequence) to be invariably conserved in all aligned sequences, including the newly predicted polypeptides (Figure 38). This residue's functional/structural importance is further highlighted by its conservation among all UniProtKB/SwissProt entries manually curated to belong to the OST4 family (data not shown); these sequences originate from diverse eukaryotes ranging from amoebozoa (*Dictyostelium discoideum*) and fungi (e.g. *S. cerevisiae*, *Schizosaccharomyces pombe*) to green plants (*Oryza sativa* and *Arabidopsis thaliana*) and animals (e.g. *Drosophila melanogaster*, mouse and human). This residue was observed to induce a kink in both the yeast and human Ost4 structure solved by solution NMR (Gayen and Kang, 2011; Zubkov et al., 2004). We performed geometry analysis of the Ost4p subunit in the recently determined yeast OST complex (PDB ID: 6EZN) revealing that, in the context of the OST complex, this helix adopts a linear conformation. Further work is necessary in order to elucidate whether this evolutionarily conserved glycine residue has a biological role during the assembly of the OST complex. Interestingly, in a preliminary survey of the yeast OST complex structure, this glycine residue is observed to participate in non-bonded contacts with a valine residue (VAL463) located in TM helix 13 of the Stt3 subunit. Another conserved position of Ost4 subunits corresponds to a histidine residue (HIS26 in the yeast sequence; Figure 38); this residue is also invariant in all members of the UniProt curated OST4 family members, and participates in non-bonded contacts with residues in TM helix 3 (SER141), and TM helix 12 (SER422, PHE425 and ASP426) of yeast Stt3.

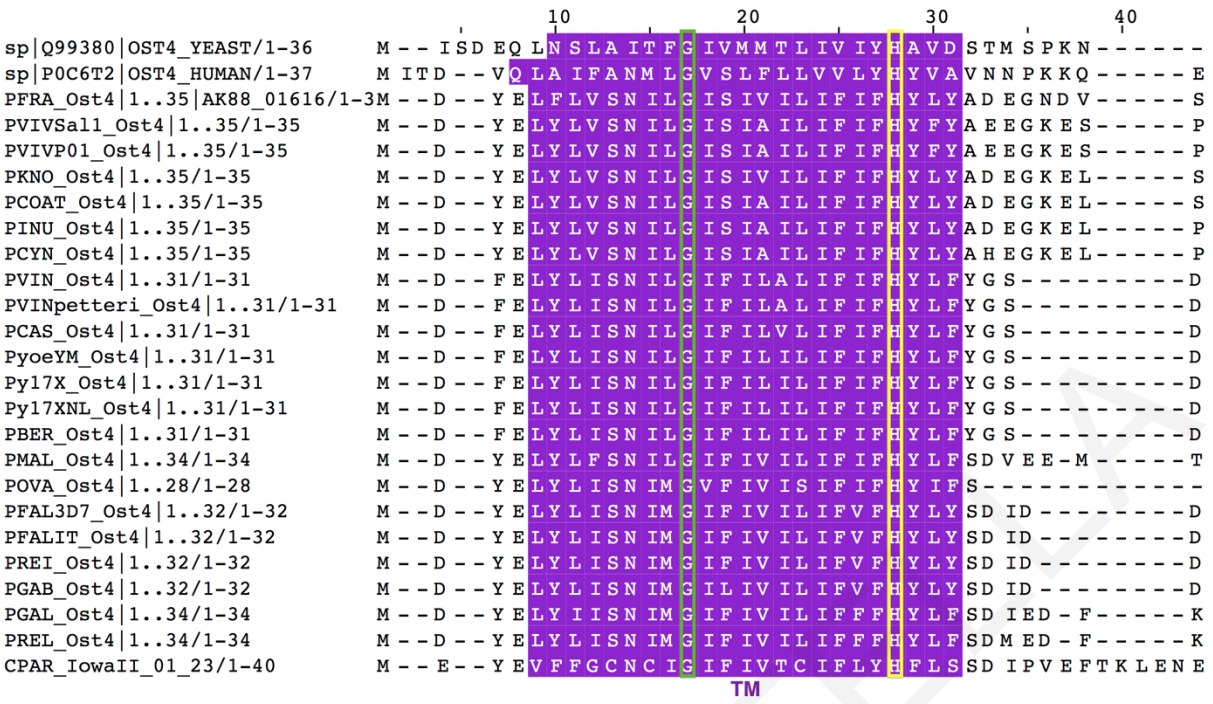



**Figure 38:** Multiple sequence alignment of the plasmodial Ost4 sequences with their predicted homolog in C. parvum (strain Iowa II) and their homologs in S. cerevisiae (UniProt AC: Q99380) and human (UniProt AC: P0C6T2). The single predicted α-helical transmembrane region (highlighted) of the protist subunits is almost perfectly aligned with those deduced from the available 3D structures for yeast (PDB ID: 6EZN; Ost4p in the OST complex) and human (PDB ID: 2LAT; Ost4p in isolation) despite their low pairwise sequence identities. The conserved glycine and histidine residues within the transmembrane region discussed in the main text are highlighted with a green and yellow box respectively.

*Ost3/Ost6*: The newly characterized Ost3/Ost6 subunits in *Plasmodium* show a varying degree of pairwise sequence conservation (from approximately 35% to 100% identity), with varying lengths (417-528 residues). Their predicted counterpart in *C. parvum* shows a marginal sequence identity (below 30% identities) with extensive insertions and deletions, resulting in 'patchy' alignments. However, the C-terminal domains are aligned with the positions of the predicted TM regions in good agreement with existing structural data (Figure 39).

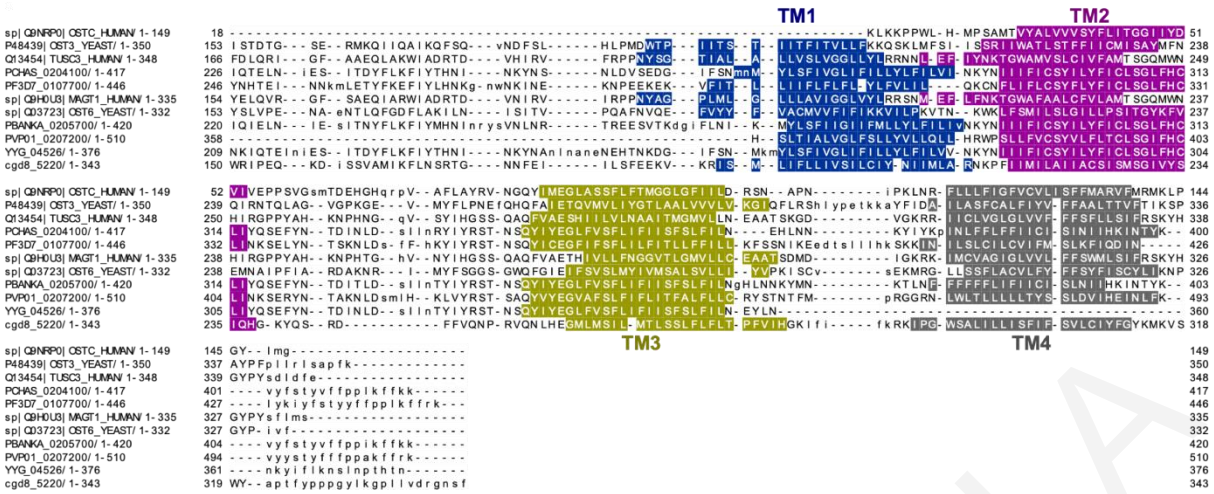



**Figure 39:** Multiple sequence alignment of ost3p/ost6p homologs against the PFAM profile HMM (PF04756.13) using the hmmalign tool of the HMMER3 package. The human OSTC homolog (DC2) is also included. Only the C-terminal transmembrane domains are displayed with the transmembrane α-helices highlighted in colour. Transmembrane segments 2-4 for OST3_YEAST were defined from the recently determined 3D structure (PDB ID: 6EZN) using the TMDET algorithm **(Tusnády et al., 2005)**. All other TM segment locations were either retrieved from the respective UniProtKB/SwissProt entries or predicted by TMHMM 2.0. Residues depicted in lowercase correspond to assignments in the model's insert/delete states (i.e. they are practically unaligned). The N-terminal part of the sequences shows very weak similarity (data not shown).

None of the *Plasmodium* Ost3/Ost6 homologs clearly matches a thioredoxin domain, even when searching with sensitive profile searches; additionally, they do not possess the CXXC motif, typical of other characterized members of the Ost3p/Ost6p family. This observation does not directly rule out the possibility that the sequences characterized in this work are genuine functional analogs of Ost3/Ost6. On the one hand, mutation of the CXXC motif does not completely abolish glycosylation in yeast (Mohd Yusuf et al., 2013). In fact, all plasmodial species with genomes available in PlasmoDB have at least a dozen predicted protein sequences annotated with the thioredoxin signature PFAM domain (PFAM: PF00085). Additionally, previous works have specifically reported oxidoreductase-independent roles for Ost3p/Ost6p in N-glycosylation (Mohd Yusuf et al., 2013; Shrimal et al., 2017), while protein families with the thioredoxin fold maintain catalytic activity in the absence of the CXXC motif (see (Atkinson and Babbitt, 2009) references therein). Also, plant Ost3 homologs have been predicted to have a thioredoxin fold but without the CXXC motif (Farid et al., 2013), an interesting finding from both a functional and an evolutionary perspective. However, another plausible explanation is that other thioredoxin-like proteins from *Plasmodium* could compensate for the missing Trx-domain of Ost3.

In the recent report of the *P. berghei* Pb51 protein (Wang et al., 2017), the authors observed the conservation of the C-terminal OST3-OST6 domain in *Plasmodium* homologs. Based on their experimental data and the presence of a signal peptide and a PEXEL motif they did not further discuss the possibility that this protein is part of the OST complex. However, in our analysis, a PEXEL

motif is predicted only in a subset of plasmodial sequences with the OST3_OST6 domain (data not shown). In the light of an increasing number of eukaryotic proteins functioning in more than one unrelated biological processes or cellular compartments (Katsani et al., 2014; Ribeiro et al., 2018), plasmodial Ost3/Ost6 might be new additions in the existing catalog of moonlighting proteins.

*Ost5*: The amino acid sequences for newly characterized *Plasmodium* Ost5p subunits share significant pairwise sequence similarity to each other (40%-100% sequence identity) as expected, whereas the *C. parvum* Ost5 exhibits very low sequence identity to *Plasmodium* (<10% identities), human (~13%) and yeast (14%) homologs. This weak sequence conservation explains why such homologs can be easily missed when searching by pairwise sequence comparison methods and highlights the importance of sequence profile-based searches. Nevertheless, the generated multiple sequence alignment (Figure 40) consistently aligns the predicted TM helices against those from the experimentally determined yeast Ost5 structure.

One difference between all sequences examined in this work and the yeast homolog is the considerably longer loop connecting the predicted TM regions compared to the loop in yeast. However, more work is necessary in order to exclude the possibility of errors due to sequence-based TM helix prediction or the assignment of TM regions based on the 3D structure with TMDET. In fact, we have reasons to believe that this discrepancy might be an artifact, since sequence-based prediction on the yeast sequence results in better agreement to the rest of the predictions (data not shown).

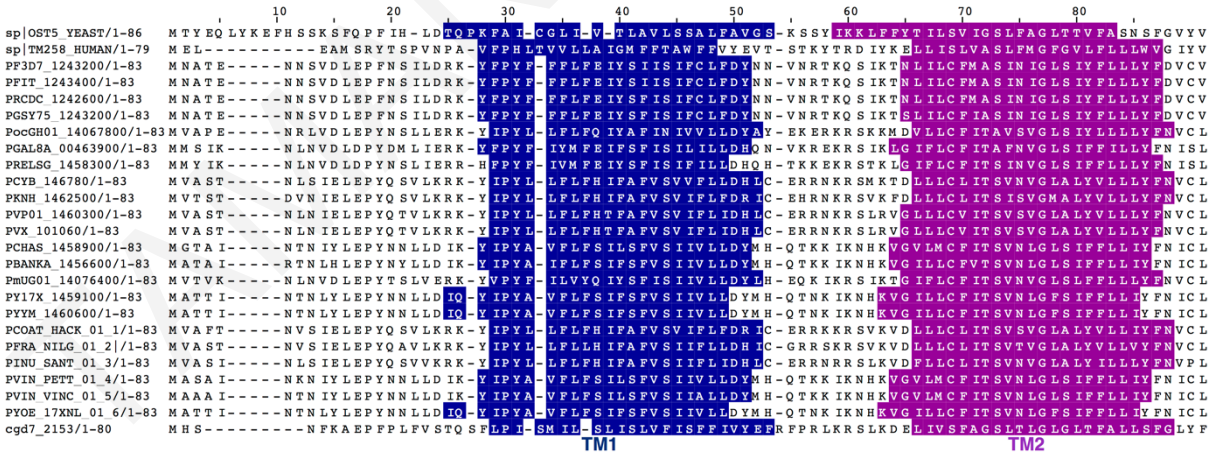



**Figure 40:** Multiple sequence alignment of the newly characterized Ost5p sequences from *Plasmodium a*nd *Cryptosporidium parvum* (strain Iowa II) with their homologs in yeast (OST5_YEAST) and human (TM258_HUMAN). Highlighted are the two transmembrane regions, defined with a similar approach to those in Ost3p/Ost6p homologs (**Figure 39**). The core of the TM segments is well aligned, while variability is observed in the length of the cytoplasmic loop connecting TM1 and TM2, with lengths ranging between 4 (yeast) to 16 amino acid residues (human). The relatively high occurrence of positively charged residues in this loop is in line with the positive inside rule (Elazar et al., 2016; Krogh et al., 2001). The observed discrepancy between the location of TM helices in yeast (experimentally determined) and those predicted in the remaining sequences disappears when

TMHMM 2.0 is applied to the yeast Ost5p sequence.

**Concluding remarks**

In this section we challenged the currently established notion that only 4 out of 8 subunits of the OST complex are present in *Plasmodium* species and other selected protists. Specifically, for each of these four "missing" OST subunits, we searched in the literature and in specialized databases for evidence supporting and validating our findings through syntenic neighborhood conservation, expression data, protein domains and sequence alignments.

We provide unequivocal evidence that, with the exception of Swp1, all components of the OST complex can be reliably identified within completely sequenced plasmodial genomes. In fact, most of the subunits currently considered as absent refer to uncharacterized protein sequences already existing in the sequence databases. Furthermore, the main reason why the unusually short Ost4 subunit has not been identified so far is the failure of gene-prediction pipelines to detect such short coding sequences.

## 3.5. Sequence and Structural Signatures of CBRs

This study aims to designate structural signatures of CBR-rich protein sequences based on data retrieved from experimentally X-RAY solved 3D protein structures. Our starting point is the analysis of secondary structure elements and accessible surface areas of CBRs in 3D structures. We used a carefully compiled dataset of non-redundant protein chains retrieved from Protein Data Bank (PDB; (Berman et al., 2000)), to represent proteins with known structures. CAST was employed in order to detect CBRs and at the same time retaining information about the CBR regions and the associated overrepresented residue types. PDB structures were analyzed by NACCESS and DSSP in order to obtain for each amino acid residue the Relative Accessible Surface Area (RASA; (Hubbard and Thornton, 1993)) and secondary structure content (DSSP secondary structure assignments (W. G. Kabsch and Sander, 1983)), respectively. For each CBR detected by CAST we extracted RASA and secondary structure information per residue, using their normalized content to generate a structure-based vector for representing individual CBRs. An example mapping a Q-rich region to its secondary structure pattern is shown in **Figure 41**.

```
>3ITFA 145 XRAY  1.450  0.169  0.189 no Periplasmic protein CpxP <CPXP_ECOLI(21-150)
[Escherichia coli]
Original: HHHHHHSSGLVPRGSAAEVGSGDNWHPGEELTQRSTQSHMFDGISLTEHQRQQMRDLMQQARHEQPPVNVSELET
CAST:     XXXXXXSSGLVPRGSAAEVGSGDNWHPGEELTXRSTXSHMFDGISLTEHXRXXMRDLMXXARHEXPPVNVSELET
DSSP:                                          STTTT   HHHHHHHHHHHHHHHHHS     HHHHHH

Original: MHRLVTAENFDENAVRAQAEKMANEQIARQVEMAKVRNQMYRLLTPEQQAVLNEKHQQRMEQLRDVTQWQ
CAST:     MHRLVTAENFDENAVRAXAEKMANEXIARXVEMAKVRNXMYRLLTPEXXAVLNEKHXXRMEXLRDVTXWX
DSSP:     HHHHHT SS  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTS HHHHHHHHHHHHHHHHTT GGGG
```

**Figure 41:** Cartoon figure of 3ITF_A protein sequence and secondary structure. The Q-rich CBR is colored red both in unmasked and masked FASTA sequence while, with green we mark the His-tag which is not taken in consideration is subsequent analyses.

**General summary of CBRs**

Our initial non-redundant dataset was composed of 4424 protein structures, where 83 protein structures were excluded from further analysis due to problematic PDB entries (See Data and Methods). Thus, the final non-redundant dataset was composed of 4341 protein structures and sequences (Table 27). Of those 4341 proteins, 1290 proteins (29,72%) contained at least one CBR detected by CAST mode 25. Even though, the total number of experimentally solved structures grows exponentially in PDB, the number of CBRs in experimentally solved protein structures remains quite small. This is not surprising, since earlier studies showed that PDB is heavily biased against CBR-rich proteins (Berman et al., 2000; Huntley and Golding, 2002; Tamana et al., 2012; Wootton and Federhen, 1993) presumably due to the induced difficulty of biased regions to crystalize (Bannen et al., 2007; Coletta et al., 2010).

**Table 27:** General summary of the protein structures dataset. The **SM(x)** column denotes the total number of masked residues, **SN(y)** denotes the total number of residues in our dataset and **SNyPmasked** is the total number of residues in proteins with at least one CBR in this dataset. All statistics were calculated based on the total number of proteins left after the elimination of 83 proteins with problematic PDB files (see Methods for details).

| Initial #sequences | #sequences post-processing | #sequences with CBRs | #CBRs | SM(x) | SN(y) | SNyPmasked |
|---|---|---|---|---|---|---|
| 4424 | 4341 | 1290 | 1634 | 17487 | 1021731 | 403241 |

Comparing the global composition of residues in our non-redundant PDB structures dataset, we observed that the most abundant residue types are the hydrophobic Ala, Gly, Leu, Val, the polar Ser and charged Glu (Figure 42). Rare residues are polar Cys and Trp, which compose approximately 1% of the global composition. These values serve as a descriptor of the average global compositional bias of the proteins in our dataset and may be a source of CBR in protein sequences.

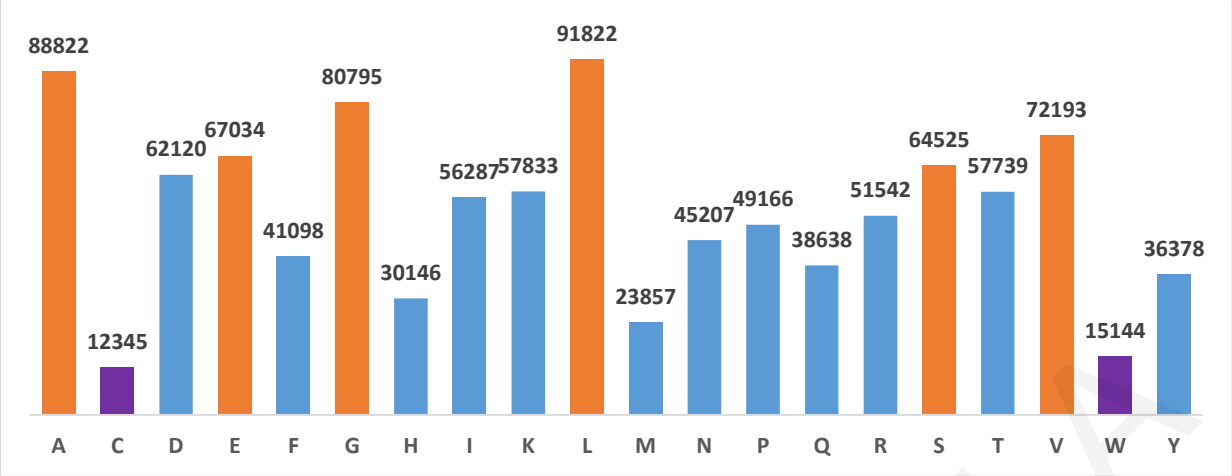



**Figure 42:** Background frequencies ($N_y \in \{A, R, ..., V\}$) for the twenty standard amino acids in the PISCES dataset. **Orange** columns signify the most abundant residue types, while **Purple** columns signify rare residue types and **Blue** the remaining residue types.

It's worth mentioning that certain residue types are detected in very low numbers by CAST mode 25 (Figure 43). Such residues include mostly the hydrophobic Ile, Leu, Met, Phe, rare types Trp, Tyr, Val, charged Arg and Asp and the polar Cys. Leu is the third less detected residue type after Ile and Phe, but it is also one of the most abundant residue types in our dataset. One possible explanation of this observation may stem from the very negative scores related to Leu in the BLOSUM62 matrix. Thus, only regions with very high composition in Leu would be expected to be detected by CAST.

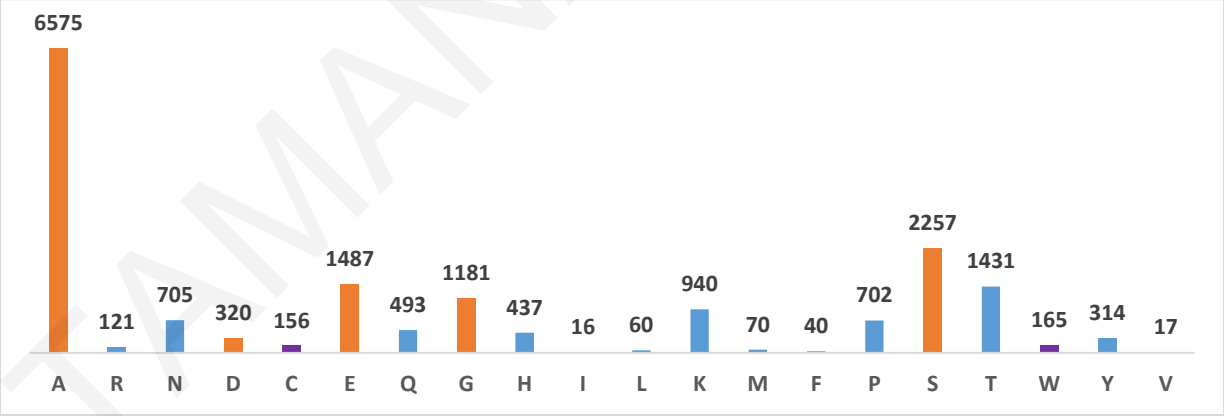



**Figure 43**: Illustrates the total number of masked amino acids (SM(x)). **Orange** columns signify the most abundant residue types and **Purple** the rare residue types in the PISCES dataset.

Notably, the local compositional bias detected does not seem to correlate with global bias (see Figure 42, Figure 43). More specifically, the total number of Ile residues is 56287 while the number of Ile residues being masked by CAST is 16. This finding is in sharp contrast with the fact that Ile is more frequent than what would be expected if residues were distributed

in a uniform fashion ($N_{Ile}$ = 5.5% compared to 5% for uniform distribution in PISCES). In BLOSUM62, Ile-Ile scoring is 4 and thus, CAST mode 25 detect Ile as a biased residue type only if there are 7 or more Ile residues in a protein sequence. Another possible explanation could be the fact that Ile scores positively only with Leu (Ile-Leu scoring is 2) and Met (Ile-Met scoring is 1) indicating that Ile-rich CBRs should also be enriched with Leu/Met among other residue types in order for CAST to detect such CBRs. Reviewing these Ile-rich CBRs, we observed that all Ile biased regions were enriched with Leu than Ile but CAST mode 25 identified Ile as the causing bias type. This observation suggests that Ile-rich CBRs could be artifact of CAST mode 25.

Another residue type that follows the same pattern is Phe. Apart from itself, (Phe-Phe scoring in BLOSUM62 is 6) Phe scores positively only against Tyr (Phe-Tyr scoring is 3) and Trp (Phe-Trp scoring is 1). Thus, a similar explanation with Ile may also apply for Phe.

Apart from the hydrophobic Ala, the smallest residue Gly, the polar Ser, Thr, charged Glu and Lys are the most common residue types being masked by CAST. Furthermore, Ala and Ser have negative scores with small absolute values and thus, their high CBR detection may be an artifact of the method.

**Structural features of CBRs**

Compositionally Biased Regions were often mistaken as 'junk' peptides due to their tendency to conform into non-globular domains or being in disorder state (Dunker et al., 2001; Romero et al., 2000; Saqi, 1995; Toll-Riera et al., 2012). Such tendency is further accompanied by difficulty in solving the 3D protein structure with commonly-used experimental procedures such as X-Ray crystallography and NMR spectroscopy (Coletta et al., 2010; Crick et al., 2006; Kumar et al., 2017; Romero et al., 2000).

However, in this study we observed approximately 30% of the protein structures to have at least one CBR. The clustering analysis using the k-means algorithm yielded into 4 distinct clusters (Table 28) based on the individual values of Relative Accessible Surface Area (RASA) and secondary structure (DSSP) patterns (see Data and Methods). Additionally, we computed descriptive statistics, Fischer's test, Kruskal-Wallis (Kruskal and Wallis, 1952; Spurrier, 2003) and Dunn's non-parametric tests (Dunn, 1964, 1961) for assessing whether statistical differences emerge between the discrete clusters. In Table 28, we noticed that most CBRs rarely conform into isolated β-bridges (B), $3_{10}$-helices (G), π-helices (I), hydrogen bonded turns (T) or bends (S) but rather show a preference into Disorder (D), Loop/Irregular

(L), Extended (E) and α-helices (H) secondary structures. Based on this observation, we choose to report the residue-type analyses by merging all helical secondary structures into one category, namely Helix (H + G + I), and the remaining patterns were merged into Extended (E + B) or Loop/Irregular (L + T + S) secondary structure patterns respectively. Thus, each structure-derived cluster represents one of the major secondary structure classes, Helix (Cluster2), Extended (Cluster3) and Loop/Irregular (Cluster4) plus Disorder (Cluster1).

**Table 28:** Descriptive statistics of the resulted k-means clusters based on the CBRs structural and sequence features analysis. **Orange:** average region complexity and Shannon entropy, **Purple:** average/median RASA, **Green:** over-represented DSSP pattern, **Blue**: average Hydrophobicity and **Red:** average Net charge. **Abbreviated columns:** PL (Protein Length), RL (Region Length), CS (CAST score), RC (Region Complexity), SE (Shannon Entropy), AH (Average Hydrophobicity), Ch (Average Charge), NetCh (Net Charge), ACC (average RAS value), D (Disorder), L (Loop/Irregular structure), H (α-helix), B (β-bridge), E (Extended strand), G ($3_{10}$-helix), I (π-helix), T (Turn), S (Bend). **Abbreviated rows:** SD (Standard Deviation) and SE (Standard Error).

| | #CBRs | ACC | D | L | H | B | E | G | I | T | S | PL | RL | CS | RC | SE | Hyd | Ch | NetCh | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CLU1** | 198 | **0.93** | **0.84** | 0.06 | 0.04 | 0.00 | 0.03 | 0.01 | 0.00 | 0.02 | 0.01 | 357.01 | 25.59 | 36.83 | **1.91** | **1.67** | **1.44** | -0.08 | **0.27** | **Average** |
| | | 0.59 | 0.39 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 21.00 | 4.00 | 25.00 | 0.00 | 0.00 | -0.42 | -1.00 | 0.00 | **Min** |
| | | 1.34 | 1.00 | 0.61 | 0.50 | 0.12 | 0.50 | 0.15 | 0.00 | 0.32 | 0.21 | 1305.00 | 239.00 | 140.00 | 3.50 | 2.66 | 3.49 | 0.45 | 1.00 | **Max** |
| | | **1.00** | **0.95** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 331.50 | 21.00 | 31.00 | 1.95 | 1.77 | **1.47** | 0.00 | **0.18** | **Median** |
| | | 0.10 | 0.19 | 0.13 | 0.09 | 0.01 | 0.07 | 0.03 | 0.00 | 0.05 | 0.03 | 208.73 | 22.91 | 16.51 | 0.73 | 0.57 | 0.74 | 0.24 | 0.25 | **SD** |
| | | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 14.83 | 1.63 | 1.17 | 0.05 | 0.04 | 0.05 | 0.02 | 0.02 | **SE** |
| **CLU2** | 551 | **0.34** | 0.02 | 0.03 | 0.08 | 0.02 | **0.48** | 0.06 | 0.00 | 0.18 | 0.14 | 325.85 | 43.87 | 31.38 | **2.23** | **1.91** | **1.16** | -0.02 | **0.19** | **Average** |
| | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 41.00 | 3.00 | 25.00 | 0.00 | 0.00 | -0.30 | -1.00 | 0.00 | **Min** |
| | | 0.89 | 0.40 | 0.47 | 0.43 | 0.38 | 1.00 | 1.00 | 0.00 | 0.89 | 1.00 | 1305.00 | 477.00 | 131.00 | 3.92 | 2.81 | 3.33 | 0.75 | 1.00 | **Max** |
| | | **0.34** | 0.00 | 0.00 | 0.00 | 0.00 | **0.48** | 0.00 | 0.00 | 0.15 | 0.11 | 286.00 | 23.00 | 28.00 | 2.28 | 2.04 | **1.01** | 0.00 | **0.15** | **Median** |
| | | 0.14 | 0.06 | 0.09 | 0.12 | 0.04 | 0.24 | 0.13 | 0.00 | 0.14 | 0.13 | 195.52 | 55.27 | 10.26 | 0.91 | 0.63 | 0.56 | 0.17 | 0.19 | **SD** |
| | | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 8.33 | 2.35 | 0.44 | 0.04 | 0.03 | 0.02 | 0.01 | 0.01 | **SE** |
| **CLU3** | 599 | **0.34** | 0.02 | 0.03 | **0.66** | 0.01 | 0.08 | 0.03 | 0.00 | 0.10 | 0.06 | 306.64 | 49.95 | 32.46 | **2.38** | **2.00** | **1.11** | -0.03 | **0.24** | **Average** |
| | | 0.01 | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 21.00 | 3.00 | 25.00 | 0.00 | 0.00 | -0.45 | -0.90 | 0.00 | **Min** |
| | | 0.99 | 0.43 | 0.50 | 1.00 | 0.18 | 0.40 | 0.71 | 0.28 | 0.75 | 0.45 | 1193.00 | 390.00 | 150.00 | 3.76 | 2.73 | 3.20 | 0.56 | 0.90 | **Max** |
| | | **0.34** | 0.00 | 0.00 | **0.62** | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.05 | 276.00 | 31.00 | 29.00 | 2.57 | 2.19 | **0.93** | 0.00 | **0.18** | **Median** |
| | | 0.15 | 0.07 | 0.09 | 0.21 | 0.02 | 0.10 | 0.07 | 0.02 | 0.10 | 0.07 | 169.58 | 55.21 | 12.06 | 0.85 | 0.59 | 0.55 | 0.15 | 0.19 | **SD** |
| | | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.93 | 2.26 | 0.49 | 0.03 | 0.02 | 0.02 | 0.01 | 0.01 | **SE** |
| **CLU4** | 286 | **0.39** | 0.04 | **0.88** | 0.04 | 0.00 | 0.02 | 0.01 | 0.00 | 0.01 | 0.01 | 324.78 | 30.66 | 31.06 | **2.11** | **1.84** | **1.22** | -0.04 | **0.23** | **Average** |
| | | 0.01 | 0.00 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.00 | 3.00 | 25.00 | 0.00 | 0.00 | -0.13 | -0.88 | 0.00 | **Min** |
| | | 1.12 | 0.47 | 1.00 | 0.46 | 0.09 | 0.38 | 0.50 | 0.00 | 0.32 | 0.18 | 1193.00 | 191.00 | 172.00 | 3.68 | 2.78 | 3.15 | 0.67 | 1.00 | **Max** |
| | | **0.38** | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 294.00 | 23.00 | 28.00 | 2.25 | 2.02 | **1.05** | 0.00 | **0.17** | **Median** |
| | | 0.18 | 0.09 | 0.17 | 0.09 | 0.01 | 0.06 | 0.04 | 0.00 | 0.03 | 0.03 | 188.91 | 27.56 | 12.26 | 0.82 | 0.60 | 0.61 | 0.19 | 0.21 | **SD** |
| | | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 11.17 | 1.63 | 0.73 | 0.05 | 0.04 | 0.04 | 0.01 | 0.01 | **SE** |

Evidently, all clusters corresponding to well-structured regions (Clusters 2-4) tend to have decreased RASA (i.e. $0 \leq$ median RASA value $\leq 0.4$) in line with their folded nature (Kruskall-Wallis test p<2.2e-16 followed by Dunn's test, p<0.001; Figure 44).

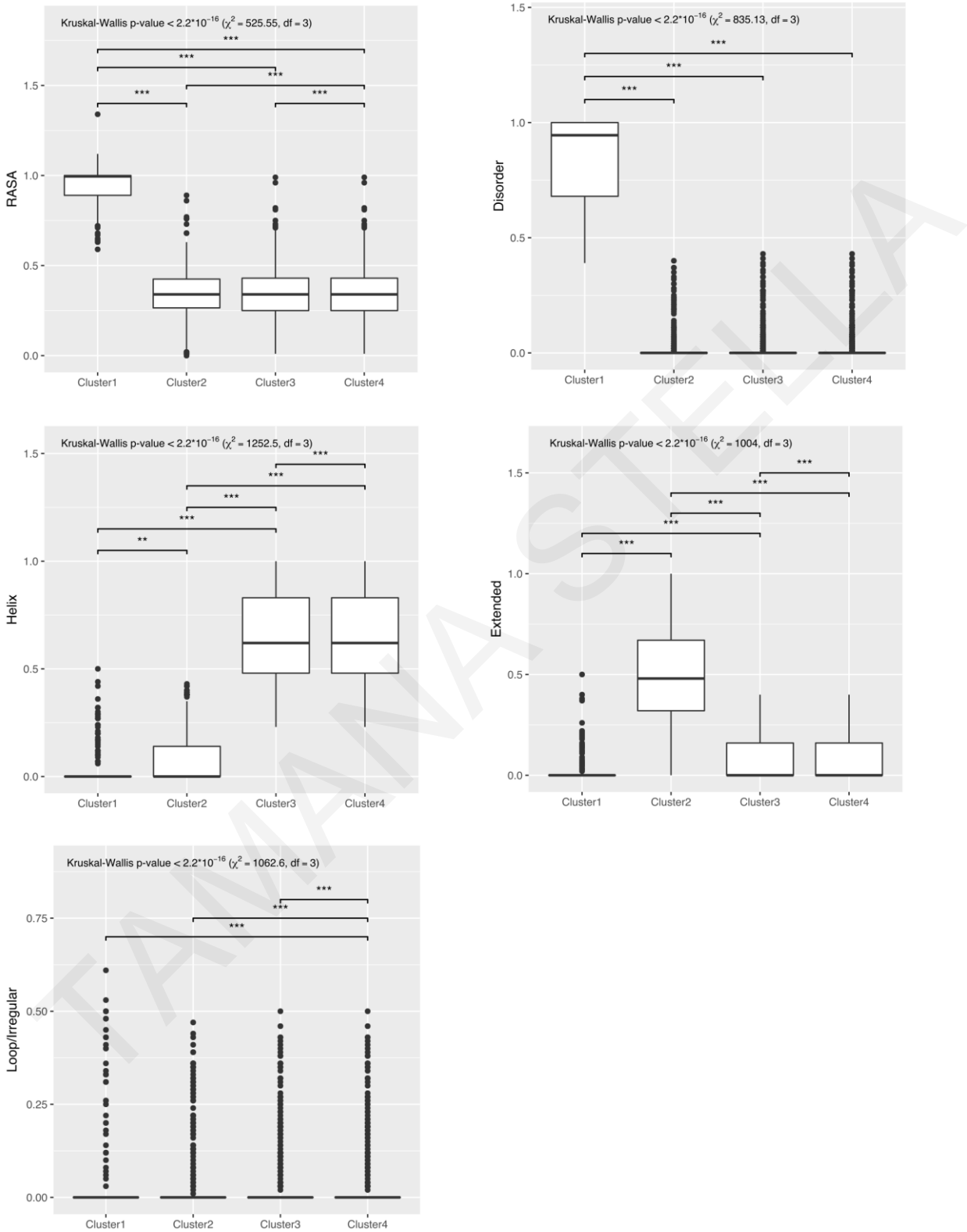



**Figure 44:** Structural differences highlighted across structure-based clusters. Boxplots depict statistically significant structural properties across different clusters: Kruskal-Wallis test followed by Dunn's post-hoc test. Only significant differences are presented ($p < 0.001$).

The clusters labeled as Buried, have average hydrophobicity and net charge of approximately 1.16 and 0.22 (Figure 45) respectively while the over-represented DSSP patterns are Extended, Helix and Loop/Irregular structure (Table 28; **Table 29**). Based on our selected hydrophobicity scale (Hessa et al., 2005), more negative hydrophobicity values indicate more hydrophobic residue types whereas, higher values indicate hydrophilic residues. During protein synthesis, the amino acids are put together into a polypeptide chain where it folds into regular local structures (e.g. helices, β-sheets and coils) (Kabsch and Sander, 1983; Lee and Richards, 1971). Each of the twenty different amino acids have specific characteristics defined by the side chain, which provides it with its unique role in a protein structure (Wampler, 2010). Based on the propensity of the side chain to be in contact with polar solvent like water, hydrophobic amino acids build-up the hydrophobic core, which is not accessible to solvent, while polar or charged residue types tend to build-up the polar surface in order to be in contact with the environment (Wampler, 2010). Thus, indicating that the secondary structure pattern in which a CBR may conform (i.e. structured or disordered) depends both on the physicochemical properties of the causing bias residue type(s) and their interactions with the remaining residues of the region to form stable polypeptide chains (Figure 45).

Remarkably, Cluster1 is the only cluster with high median RASA value labeled as Accessible (median RASA = 1.0), highest median average hydrophobicity value (1.47) and median average Net charge (0.18) while the over-represented secondary pattern is Disorder ( Table 28; **Table 29**). Although, the number of CBRs in disorder state varies greatly (min = 0.39 and max=1.0) between cluster CBRs, median value ($\alpha$ = 0.97) indicate that disorder is the prevalence state. We should note, however, we treated missing structural residues as disorder and highly accessible based mostly on the intuition that it won't be any missing residues from the hydrophobic core (i.e. it should be packed in order to keep the whole protein together).

Nonetheless, this finding suggests that CBRs grouped in Cluster1 are composed mostly by charged (i.e. present energetically favorable contacts with water) residue types, such as Aspartic acid and Glutamic acid, which preferentially are exposed to solvent as to form hydrogen bonds. Consequently, the excess of charge in these CBRs combined with the high average hydrophobicity value may induce an unordered state for the region (Wampler, 2010; Zhang et al., 2007). We tested this hypothesis by computing the Kruskal-Wallis test (Kruskal and Wallis, 1952; Spurrier, 2003) followed by the post-hoc Dunn's test

(Dunn, 1964, 1961) on per cluster's average hydrophobicity and net charge values (Figure 45). At 3 degrees of freedom and p-value equal to $9.496e^{-09}$ (hydrophobicity) and 0.0001801 (net charge), the Kruskal-Wallis null hypothesis is rejected while, the Dunn's post-hoc analysis indicated that the statistically significant differences are between Cluster1 and the remaining three clusters. Hence, higher CBR hydrophobicity value promotes disorder.

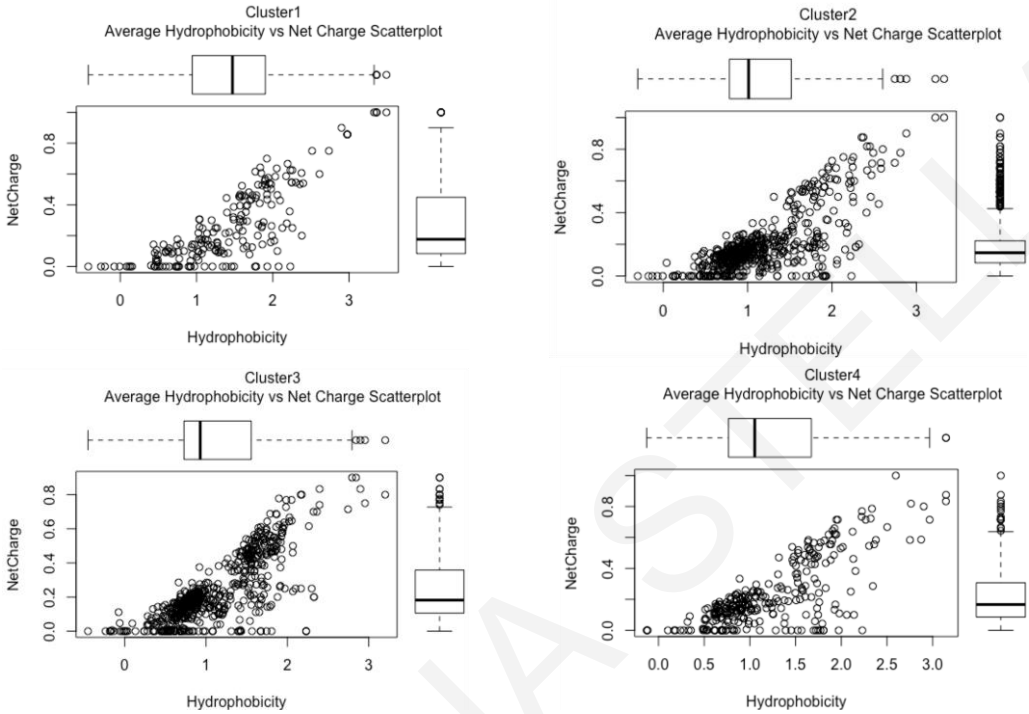



**Figure 45:** Average hydrophobicity (bottom x-axis) versus mean Net charge (left y-axis) per cluster-based scatter plot of CBRs. Per cluster boxplot: **(i)** top x-axis: mean hydrophobicity and **(ii)** right y-axis: mean net change.

Reviewing Fischers' test results, we observed that the statistically significant over-represented residue types in Cluster1, apart from the negatively charged Aspartic acid, are Glycine and Proline (**Table 29**; **Figure 46**). In most cases, Glycine is found at the surface of proteins, within loop- or coil (lack of secondary structure) regions, providing high flexibility to the polypeptide chain (Wampler, 2010). In contrast, Proline is mostly found buried inside the protein and this is unexpected as Proline provides rigidity to the polypeptide chain due to its rotationally constrained rigid-ring structure (Bhagavan and Ha, 2011). However, Campen and colleagues indicated, Proline as one of the most disorder-promoting residues despite the non-polar nature of its side chains (Campen et al., 2008; Theillet et al., 2013). Additionally, a number of research groups addressing the "disorder-promoting residue types" illustrated that along with Aspartic acid and Proline, Methionine, Lysine, Arginine, Serine, Glutamic acid and Glutamine are the most

commonly “disorder-promoting residue types” (Campen et al., 2008; Theillet et al., 2013; Williams et al., 2001).

Cluster4 is the only cluster that Fischers’ test did not provide which residue type is over-/under-represented in Loop/Irregular secondary structures when in CBRs, suggesting that all CBRs may conform into that particular secondary structure (**Table 29**).

**Table 29:** A summary table of the results from the sequence and structural features analysis. **Red** denote the **structural features** and **statistically significant over-represented** residue types in the individual clusters. **#CBRs:** number of CBRs; **PAL:** Protein Average Length; **ACL:** Average Region Length; **AH:** Average Hydrophobicity; **AC:** Average Charge; **ANC:** Average Net Charge; **AA:** Average Accessibility; **SS:** Secondary Structure; **SAAs:** Significant Amino Acids.

| CID | #CBRs | PAL | ARL | AH | AC | ANC | AA | SS | SAAs |
|---|---|---|---|---|---|---|---|---|---|
| **Cluster1** | 198 | 357 | 25.59 | 1.44 | -0.08 | 0.27 | **Exposed** (0.93) | **Disorder** | **D, G, P** |
| **Cluster2** | 551 | 325.85 | 43.87 | 1.16 | -0.02 | 0.19 | **Buried** (0.34) | **Extended** | **N, G, S, T** |
| **Cluster3** | 599 | 306.64 | 49.95 | 1.11 | -0.03 | 0.24 | **Buried** (0.34) | **Helix** | **A, E** |
| **Cluster4** | 286 | 324.78 | 30.66 | 1.22 | -0.04 | 0.23 | **Buried** (0.39) | **Loop/Irregular structure** | - |

Fischer’s test demonstrated that over-represented residue types exist within clusters: D/G/P (Cluster 1), A/E (Cluster 2) and G/N/S/T (Cluster 3) (**Table 29**). With the exception of Ile (n=4) different CBR types seem to form 3 distinct clusters based on average structural properties (Table 30).

**Table 30:** Average accessible surface area and structural pattern values of the x-rich CBRS. **Green:** Fischer’s’ test significant over-represented residues. **Orange:** Fischer’s’ test significant under-represented residues. **Purple:** the highest average RASA and DSSP values.

| AA | RASA | Disorder | Loop | Helix | Extended | AA | RASA | Disorder | Loop | Helix | Extended |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.35 | 0.08 | 0.16 | 0.39 | 0.15 | M | 0.47 | 0.19 | 0.16 | 0.29 | 0.16 |
| C | 0.27 | 0.04 | 0.15 | 0.28 | 0.13 | N | 0.42 | 0.09 | 0.17 | 0.16 | 0.30 |
| D | 0.65 | 0.32 | 0.19 | 0.12 | 0.18 | P | 0.62 | 0.22 | 0.23 | 0.17 | 0.21 |
| E | 0.51 | 0.14 | 0.20 | 0.41 | 0.10 | Q | 0.50 | 0.18 | 0.22 | 0.35 | 0.11 |
| F | 0.35 | 0.07 | 0.22 | 0.35 | 0.07 | R | 0.47 | 0.12 | 0.15 | 0.38 | 0.19 |
| G | 0.40 | 0.17 | 0.13 | 0.22 | 0.22 | S | 0.45 | 0.15 | 0.18 | 0.18 | 0.26 |
| H | 0.36 | 0.12 | 0.13 | 0.24 | 0.25 | T | 0.39 | 0.04 | 0.23 | 0.15 | 0.33 |
| I | 0.74 | 0.50 | 0.00 | 0.50 | 0.00 | V | 0.22 | 0.00 | 0.20 | 0.45 | 0.14 |
| K | 0.50 | 0.13 | 0.17 | 0.33 | 0.17 | W | 0.18 | 0.00 | 0.26 | 0.25 | 0.23 |
| L | 0.35 | 0.25 | 0.00 | 0.47 | 0.15 | Y | 0.20 | 0.01 | 0.25 | 0.25 | 0.27 |

Although, CAST mode 25 detected a very small number of Ile-rich CBRs to be statistically verified, this notion is consistent with previous studies addressing the order/disorder-promoting residue types (Campen et al., 2008; Theillet et al., 2013; Williams et al., 2001; Zhang et al., 2007). **Figure 46** summarizes this information, suggesting that no CBR types are associated with uniformly high potential for intrinsic disorder. This is particularly

interesting as the currently established notion is that CBRs favored Disorder domains (Kumari et al., 2015; Peng et al., 2015).

Thus, additional sequence features should be explored for predicting the disorderliness of particular CBRs.

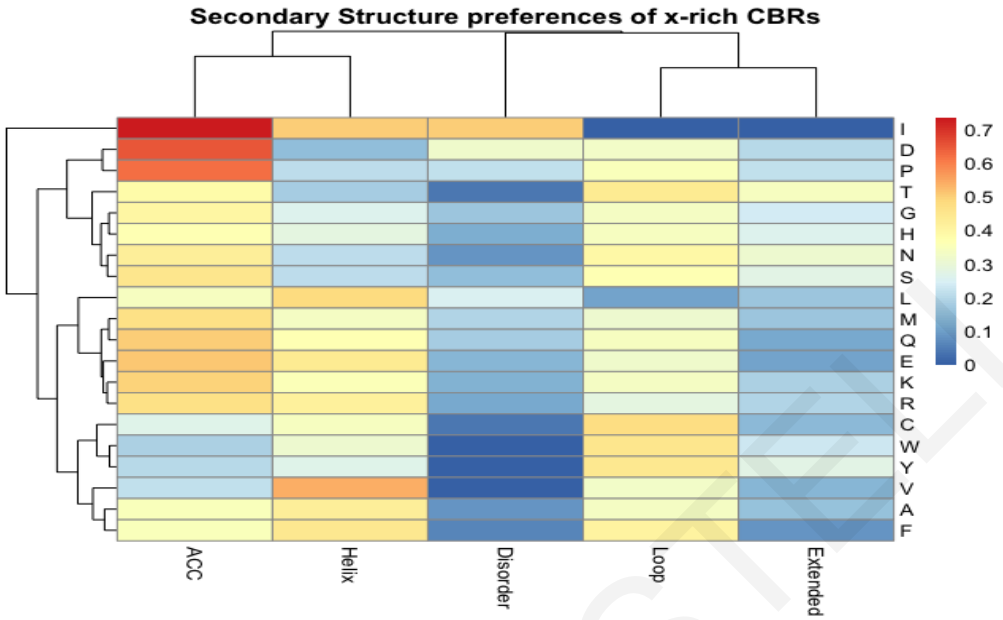



**Figure 46:** A heatmap plot depicting CBRs structural pattern preferences based on a presence (1)/absence (-1) pattern. Each column illustrates one of the DSSP secondary structure pattern plus average accessibility (RASA). The heatmap was constructed using R package pheatmap (MRAN, 2018; RStudio Team, 2015).

**Mapping Sequence Complexity to Structural features**

In order to capture each CBRs complexity state, along with the structural features of CBRs we computed a local complexity measure (SEG complexity measure $K_1$) and Shannon entropy (SEG complexity measure $K_2$; see Data and Methods). These two measures will help us understand how a region's complexity affects its exposed/buried and secondary structure patterns. Importantly, along with the SEG-like measures and CAST's ability to detect which is the causing bias residue type in each CBR we could determine which sequence features promote well-structured regions instead of disorder.

A Box-and-Whiskers plot (also known as boxplot; Figure 47) was constructed illustrating the causing biased residue types detected by CAST mode 25 and their descriptive statistics if these regions were detected by SEG.

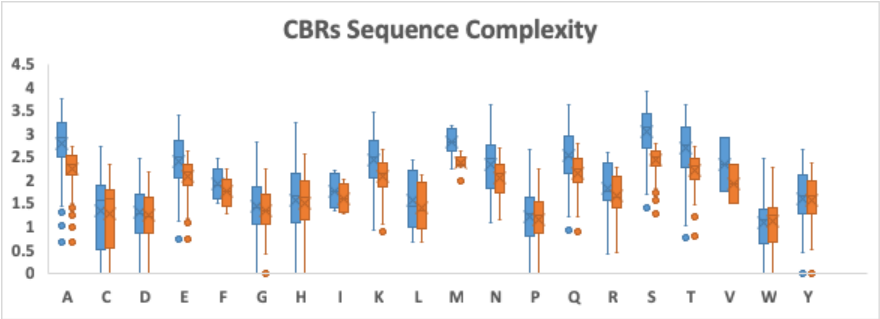



**Figure 47:** Box-and-Whiskers plot of x-rich CBRs, **SEG-like** complexity and **Shannon entropy**.

Notably, we observed that certain CBRs would not be detected as biased segments by SEG (Figure 47) as both region complexity and Shannon entropy is higher than the cutoff values set by SEG (using SEG default parameter set). Such residue types are Alanine, Glutamic acid, Serine and Threonine where their median value fall in $K_1/K_2$ values higher than $K_1/K_2 \geq 2.5$. CBRs that will also be detected by SEG would probably be enriched by Cysteine, Aspartic acid, Phenylalanine, Glycine, Histidine, Proline, Tryptophan and Tyrosine residues. In addition, when plotting the Shannon entropy value of each CAST detected CBR versus CAST normalized score (normalized by region's length; Figure 48), we observed a strong correlation ($R^2 = 0.8$) in all four clusters where many points crowding the bottom-right corner (i.e. high entropy and low normalized CAST score). The observed correlation between these measures along with structure-based clustering suggests that sequence features could highlight when CBRs will conform into well-structured regions or be in disorder state depending on their Shannon entropy value.

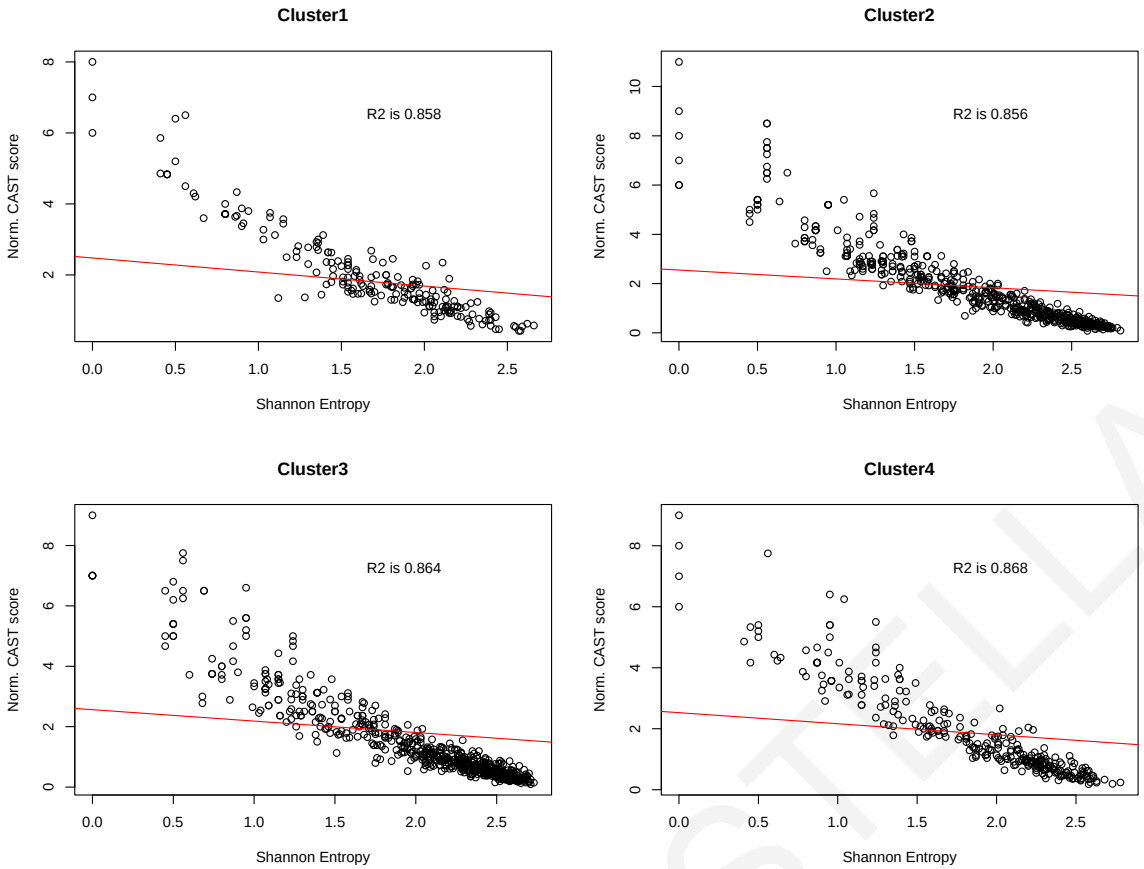



**Figure 48:** A scatterplot depicting CAST normalized score versus Shannon entropy based on the four structure-based clusters. Red line: linear-fitted regression line, x-axis: Shannon entropy and y-axis: CAST score normalized by region's length.

Following this trend, we further sub-cluster the structure-derived clusters by performing k-means clustering using sequence-derived features (Shannon entropy, average hydrophobicity, average charge and net charge). We followed a similar methodology as in structure-based clustering where we first, determined the optimal k value for each cluster (Figure 11) followed by computing the k-means algorithm in each cluster's sequence feature. In these sets of clustering we selected to use only Shannon entropy (along with mean hydrophobicity, average charge and net charge) mainly because of the strong correlation between Shannon entropy and SEG-like complexity (Figure 49).

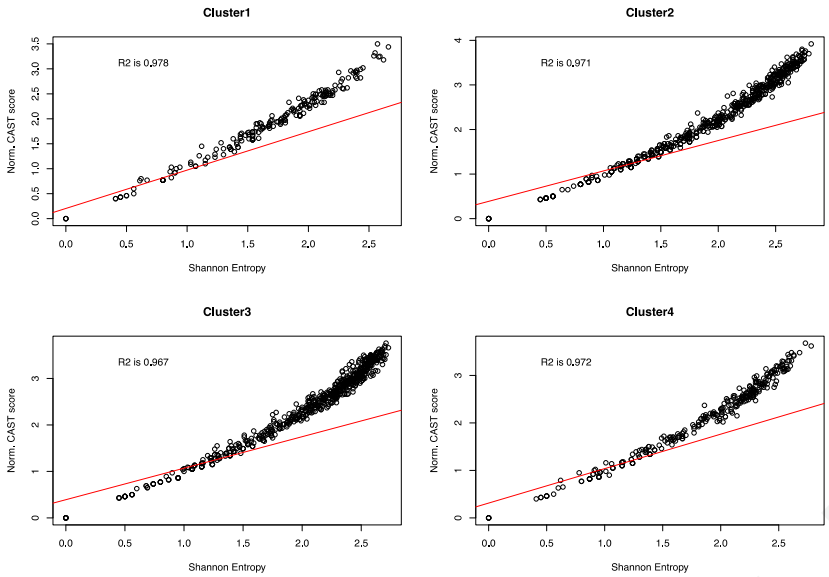



**Figure 49:** Shannon entropy versus local complexity scatterplots of the structure-derived clusters.

Sequence-derived features sub-clustering indicate that within each structure-based cluster there exist groupings corresponding to residues of specific properties (**Table 31**). The full summary table with descriptive statistics of the sequence features sub-clustering can be seen in Appendix IV. For example, Cluster1 is mostly represented by highly accessible and charged CBRs, as we noted previously, sequence features analysis indicates that higher information content CBRs (i.e. Shannon entropy > 1.3) create clusters composed of larger CBRs sharing similar hydrophobicity and charge values (e.g. sub-cluster2 and 4). Pairwise matrix plots were constructed for each structural-based cluster highlighting both their CBRs structural and sequence features. An example of such matrix plot is Figure 50 where, one can compare Cluster1 structural features against sequence features at once. All pairwise matrix plots are provided as Supplementary Material.

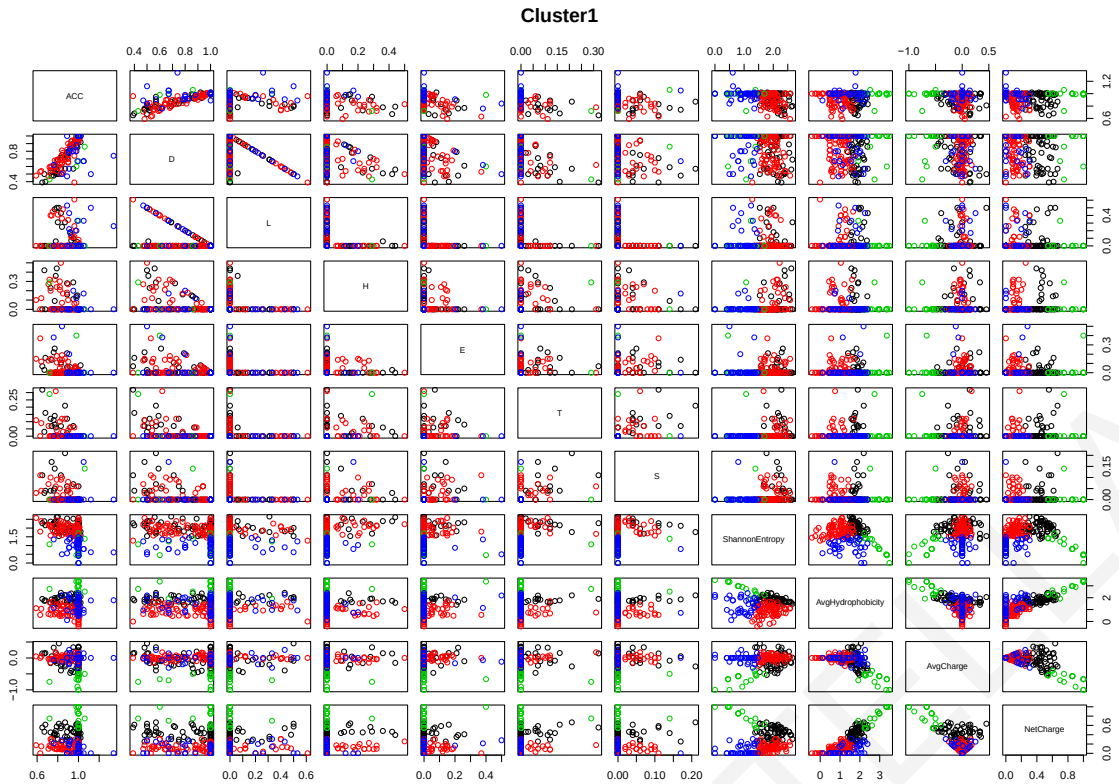



**Figure 50:** Example matrix plot of Cluster1 sub-clustering of x-rich CBRs structural and sequence features.

All clusters corresponding to well-structured regions (Clusters 2-4) illustrate scalar sub-clustering of the sequence features where lower Shannon entropy indicates medium-to-higher mean hydrophobicity and net charge but also, less accessible CBRs (**Table 31**). Although, in all cases of both structural and sequence CBRs features there are overlaps, Kruskal-Wallis and post-hoc Dunn's tests illustrated that statistically significant differences exist between each k-means sub-clustering (Table 32). Take for instance Cluster3, which was labeled as Helix due to the over-representing helical pattern, Kruskal-Wallis test of the sequence-based clustering rejected the null hypothesis of the helical patterns for the sub-cluster 6 (df = 5 and p-value < 2.2e-16). Dunn's test revealed that the observed differences are between sub-cluster 6 and sub-cluster3 (p-value = $8.8e^{-11}$) and 4 (p-value = $6.4e^{-16}$) respectively, indicating that CBRs with different sequence signatures grouped between these clusters (Table 32). Evidently, the clusters composed mostly by disorder or irregular structures (cluster1 and 4 respectively), observed statistically significant differences are found mostly regarding their buried/exposed pattern suggesting that RASA is, also, an important discriminating factor.

Interestingly, in all four structure-derived clusters there is a sub-cluster composed only by Aspartic acid and Glutamic acid rich CBRs (**Table 31**). Numerous studies have shown that Aspartic acid and Glutamic acid repeats have important biological roles owing to their

negative charges and underlying properties to interact with metal ions (Chou and Wang, 2015). For instance, Glutamic acid rich proteins serve as markers for the diagnosis of malaria (Kattenberg et al., 2012) and babesia (Mousa et al., 2013) while, Aspartic acid rich proteins are major components of the soluble organic matrix of mollusk shells (Gotliv et al., 2005; Nudelman et al., 2006; Weiner, 1979). (Chou and Wang, 2015)) performing bioinformatics analysis of 173 Aspartic acid/Glutamic acid rich repeat structures suggested that they are unique components of disorder proteins involved in gene regulation, DNA mimicry and mRNA processing (Chou and Wang, 2015).

Intrigued by the observation that these D/E-rich CBRs construct a separate cluster regardless the structural pattern one can find them, we focused on the biological significance of these sub-clusters by extracting their primary classification from the respective PDB entry. In total, we observed 221 Aspartic acid and Glutamic acid rich CBRs in our dataset divided into 36 (Cluster1; 18 in Cluster1.3), 42 (Cluster2; 32 in Cluster2.1), 99 (Cluster3; 24 in Cluster3.1) and 44 (Cluster4; 15 in Cluster4.2). It's noteworthy that all D-/E-specific sub-clusters are composed of shorter CBRs (<25 region's length), lower information content (< 2.23 Shannon entropy) but accessible CBRs (RASA > 0.4). Most protein structures are classified as metal/sugar/protein binding, translation/chaperone/transport protein or enzymes (e.g. hydrolase, transferase, kinase or isomerase) notably, some additional classifications, such as toxins, apoptosis and cell adhesion, were observed.

We selected the human transmembrane protein occludin (PDB id: 1WPA), localized at tight junctions and its functions are complex and poorly understood (Li et al., 2005), to present as an example of summarizing our analysis of structural and sequence features of CBRs (Figure 51). Occludin serves as the perfect example because the Glutamic acid rich region is 68 residues long (spanning through half of the protein structure) highlighting the mixed secondary structure patterns a CBR can conform into while, at the same time, is relatively low information content (Shannon entropy = 1.69), negatively charged (mean region charge is -0.18) and have important biological role in the formation of the paracellular barrier and in cell signaling (Li et al., 2005).

```
>1WPAA CAE24571F5DB16D2 114 XRAY  1.500  0.243  0.258 no Occludin <OCLN_HUMAN(413-522)
[Homo sapiens]
Original: GSHMEEDWIREYPPITSDQQRQLYKRNFDTGLQEYKSLQSELDEINKELSRLDKELDDYR
CAST:     GSHMXXDWIRXYPPITSDQQRQLYKRNFDTGLQXYKSLQSXLDXINKXLSRLDKXLDDYR
DSSP:             TT TS    SHHH HHHHHHHHHH HHHHHHHHHH HHHHHHHHHHHHHHHH

Original: EESEEYMAAADEYNRLKQVKGSADYKSKKNHCKQLKSKLSHIKKMVGDYDRQKT
CAST:     XXSXXYMAAADXYNRLXQVXGSADYXSXXNHCXQLXSXLSHIXXMVGDYDRQKT
DSSP:     HHS  TTSHHHHHHH HHHHHHHHHH TSHHHHHHHH HHHHHHHHHH HHH
```

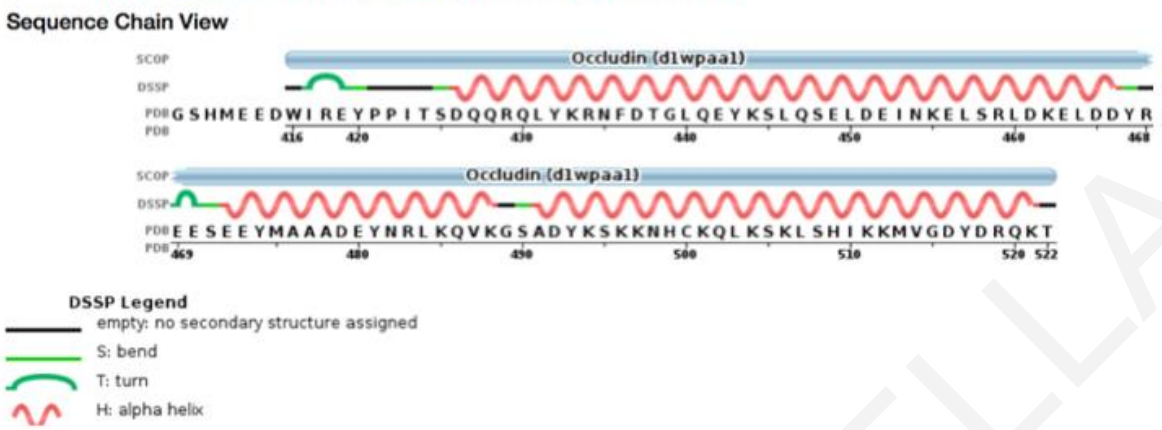






**Figure 51:** Schematic representation of human occludin (PDB id: 1WPA: A) protein structure highlighting **E**-rich CBR and secondary structure patterns. The **E**-rich CBR is colored **red** both in unmasked and masked FASTA sequence and mark the different DSSP secondary patterns found here namely **Helix** (**H**), **Bend** (**S**) and **Turn** (**T**).

The occludin structure comprises three helices that form two separate anti-parallel coiled-coils and a loop that packs tightly against one of the coiled-coils where, the Glutamic acid rich segment is found at the C-terminal distal cytoplasmic domain (Li et al., 2005). Using *in vitro* binding studies and site-directed mutagenesis, Li and colleagues, demonstrated that the highly charged and accessible surface of where the E-rich region is located, is essential for proper localization of occludin to cell–cell junctions (Li et al., 2005).

**Concluding remarks**

The last section of the results aims to designate structural signatures of CBR-rich protein sequences based on data retrieved from experimentally X-ray solved protein structures.

K-means clustering of CBRs based on their structural properties (intrinsic disorder, secondary structure and relative accessible surface area) revealed four different groups of CBRs, remarkably, corresponding to statistically significant enrichment of (i) high RASA and high disorder content, and less accessible (ii) helical, (iii) extended and (iv) loop conformations. Interestingly, most CBR types appeared in all four clusters while, Fischer's test demonstrated that over-represented residue types exist within Clusters 1-3. Further sub-clustering using sequence-derived features (Shannon entropy, average

hydrophobicity, and average charge/net charge) indicate that within each structure-based cluster there exist groupings corresponding to residues of specific properties. Interestingly, we observed D-/E-rich only CBRs to be sub-clustering together across clusters of different features and we inspected their possible biological roles. Most such protein structures are characterized to be involved in protein binding, transport processes or enzymatic activities. Notably, some additional classifications, such as toxins, apoptosis and cell adhesion, were also observed.

**Table 31:** Sub-clustering of sequence features descriptive statistics (average values) summary table. Numeric values are the median value of each feature for each of the sub-clusters. **Abbreviated columns:** CID (Cluster ID), ACC (average RAS value), D (Disorder), L (Loop/Irregular structure), H (α-helix), B (β-bridge), E (Extended strand), G ($3_{10}$-helix), I (π-helix), T (Turn), S (Bend), PL (Protein Length), RL (Region Length), CS (CAST score), RC (Region Complexity), SE (Shannon Entropy), AH (Average Hydrophobicity), Ch (Average Charge), NetCh (Net Charge), AAs (x-rich CBRs). **Purple:** highlighting D/E only clusters (in curly brackets we denote the number of CBRs observed), **Red:** Fischer's test significant over-represented x-rich CBRs.

| CID | #CBRs | ACC | D | L | H | B | E | G | I | T | S | PL | RL | CS | RC | SE | Hyd | Ch | NetCh | AAs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cluster1 – Exposed – Disorder** | | | | | | | | | | | | | | | | | | | | |
| **C1.1** | 61 | 0.99 | 0.92 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 327.00 | 25.00 | 32.00 | 2.07 | 1.83 | 1.39 | 0.00 | 0.20 | E, Q, **D**, S, **P**, N, K |
| **C1.2** | 83 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 337.00 | 20.00 | 33.00 | 1.98 | 1.80 | 1.56 | 0.00 | 0.17 | N, A, S, Q, **P**, H, F, M, T, **G**, L, I |
| **C1.3** | 18 | **1.00** | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 332.00 | **16.00** | 29.00 | 1.70 | **1.55** | **1.21** | **0.00** | **0.16** | **D {15}, E {3}** |
| **C1.4** | 36 | 0.97 | 0.86 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 322.00 | 19.50 | 30.00 | 1.93 | 1.76 | 1.48 | 0.00 | 0.21 | **P**, Q, H, N, S, **G**, C |
| **Cluster2 – Buried – Extended** | | | | | | | | | | | | | | | | | | | | |
| **C2.1** | 32 | **0.33** | 0.00 | 0.00 | 0.00 | 0.00 | **0.60** | 0.00 | 0.00 | 0.15 | 0.12 | 236.00 | **15.00** | 29.00 | 1.90 | **1.77** | **1.02** | **-0.02** | **0.13** | **D {19}, E {13}** |
| **C2.2** | 305 | 0.34 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 | 0.00 | 0.00 | 0.15 | 0.11 | 280.00 | 21.00 | 28.00 | 2.20 | 1.99 | 1.02 | 0.00 | 0.15 | M, **G**, **S**, A, Y, **N**, E, H, C, P, K, Q, **T** |
| **C2.3** | 61 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.38 | 0.00 | 0.00 | 0.16 | 0.11 | 315.00 | 22.00 | 29.00 | 2.15 | 1.89 | 0.98 | 0.00 | 0.15 | Y, **N**, K, Q, R, H, E, P |
| **C2.4** | 153 | 0.34 | 0.00 | 0.00 | 0.00 | 0.00 | 0.48 | 0.00 | 0.00 | 0.15 | 0.11 | 294.00 | 30.00 | 29.00 | 2.58 | 2.20 | 0.99 | 0.00 | 0.15 | L, F, **S**, **G**, N, Y, A, W, C, H, P, **T**, Q |
| **Cluster3 -Buried – Helix** | | | | | | | | | | | | | | | | | | | | |
| **C3.1** | 24 | **0.40** | 0.00 | 0.00 | **0.70** | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.04 | 208.00 | **20.00** | 27.50 | 2.28 | **2.07** | **1.44** | **0.00** | **0.16** | **D {8}, E {16}** |
| **C3.2** | 95 | 0.36 | 0.00 | 0.00 | 0.63 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 297.00 | 36.00 | 28.00 | 2.59 | 2.18 | 0.98 | -0.03 | 0.19 | Q, Y, T, D, H, K, N, **E**, **A** |
| **C3.3** | 60 | 0.36 | 0.00 | 0.00 | 0.68 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 259.50 | 27.50 | 30.00 | 2.43 | 2.11 | 0.96 | -0.02 | 0.21 | C, G, N, P, W, **A**, S, Q, Y, H |
| **C3.4** | 94 | 0.34 | 0.00 | 0.00 | 0.64 | 0.00 | 0.02 | 0.00 | 0.00 | 0.09 | 0.06 | 276.00 | 34.50 | 28.50 | 2.68 | 2.28 | 0.88 | 0.00 | 0.15 | F, T, I, V, Y, **A**, W, L, P, C, G |
| **C3.5** | 53 | 0.35 | 0.00 | 0.00 | 0.59 | 0.00 | 0.05 | 0.00 | 0.00 | 0.09 | 0.05 | 267.00 | 31.00 | 30.00 | 2.56 | 2.18 | 1.13 | 0.00 | 0.20 | **E**, N, R, K, Q |
| **C3.6** | 273 | 0.32 | 0.00 | 0.00 | 0.61 | 0.00 | 0.01 | 0.00 | 0.00 | 0.10 | 0.05 | 283.00 | 30.00 | 29.00 | 2.57 | 2.19 | 0.89 | 0.00 | 0.18 | M, W, **A**, C, G, N, P, F, H, V, S, Y, Q, T |
| **Cluster4 – Buried – Loop/Irregular structure** | | | | | | | | | | | | | | | | | | | | |
| **C4.1** | 45 | 0.38 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 291.00 | 26.00 | 30.00 | 2.34 | 2.06 | 0.96 | -0.01 | 0.16 | E, Q, D, S, P, N, K |

| CID | #CBRs | ACC | D | L | H | B | E | G | I | T | S | PL | RL | CS | RC | SE | Hyd | Ch | NetCh | AAs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C4.2** | 15 | **0.46** | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 268.00 | **24.00** | 31.00 | 2.29 | **2.01** | **1.31** | **0.00** | **0.19** | **D {11}, E {4}** |
| **C4.3** | 124 | 0.40 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 289.50 | 20.00 | 28.50 | 2.19 | 2.02 | 1.16 | 0.00 | 0.17 | G, N, T, W, Y, C, H, P, S, M, Q, A |
| **C4.4** | 57 | 0.38 | 0.00 | 0.94 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 306.00 | 23.00 | 28.00 | 2.20 | 2.01 | 0.95 | 0.00 | 0.14 | P, S, A, H, W, C, Y, N, T, G, F |
| **C4.5** | 19 | 0.32 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 429.00 | 23.00 | 26.00 | 2.18 | 1.96 | 0.92 | 0.00 | 0.16 | R, K, P |
| **C4.6** | 26 | 0.33 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 252.50 | 27.00 | 28.00 | 2.49 | 2.13 | 1.06 | 0.00 | 0.17 | N, H, P |

**Table 32:** Statistical tests performed for the sub-clustering analysis of the CBRs sequence features. Abbreviated columns: **KW** (Kruskal-Wallis test p-value), **C1/C2** (sub Cluster ID) and **Dp** (Dunn's test p-value adjusted for multiple comparisons). In all tests, only significant differences are presented (α=0.01).

| | ACC | | | | Disorder | | | | Helix | | | | Extended | | | | Loop | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **KW** | **C1** | **C2** | **Dp** | **KW** | **C1** | **C2** | **Dp** | **KW** | **C1** | **C2** | **Dp** | **KW** | **C1** | **C2** | **Dp** | **KW** | **C1** | **C2** | **Dp** |
| **Cluster1** | 2.41E-04 | 3 | 1 | 0.0096 | 0.04444 | - | - | - | 0.1819 | - | - | - | 0.05706 | - | - | - | 0.3588 | - | - | - |
| | | 4 | 1 | 0.0096 | | | | | | | | | | | | | | | | |
| | | 3 | 2 | 0.0096 | | | | | | | | | | | | | | | | |
| | | 4 | 2 | 0.0085 | | | | | | | | | | | | | | | | |
| **Cluster2** | 2.64E-14 | 2 | 1 | 3.00E-06 | 1.55E-05 | 4 | 2 | 3.40E-05 | < 2.20E-16 | 2.2 | 2.1 | 0.0006 | 0.005693 | - | - | - | 2.22E-12 | 2.2 | 2.1 | 0.00059 |
| | | 4 | 1 | 7.50E-10 | | | | | | 2.4 | 2.2 | < 2.00E-16 | | | | | | 2.3 | 2.2 | 4.60E-05 |
| | | 3 | 2 | 3.60E-05 | | | | | | 2.4 | 2.3 | 2.40E-06 | | | | | | 2.4 | 2.2 | 4.20E-10 |
| | | 4 | 3 | 2.20E-09 | | | | | | | | | | | | | | | | |
| **Cluster3** | < 2.20E-16 | 4 | 1 | 4.70E-11 | 0.0008285 | - | - | - | <2.20E-16 | 6 | 3 | 8.80E-11 | <2.20E-16 | 5 | 1 | < 2.00E-16 | 6.76E-06 | 6 | 4 | 2.20E+00 |
| | | 6 | 1 | 4.90E-06 | | | | | | 6 | 4 | 6.40E-16 | | 5 | 2 | < 2.00E-16 | | | | |
| | | 4 | 2 | 2.00E-16 | | | | | | | | | | 5 | 3 | < 2.00E-16 | | | | |
| | | 6 | 2 | 3.40E-14 | | | | | | | | | | 5 | 4 | < 2.00E-16 | | | | |
| | | 4 | 3 | 2.20E-09 | | | | | | | | | | 6 | 1 | 0.00011 | | | | |
| | | 5 | 4 | 2.00E-16 | | | | | | | | | | 6 | 2 | 0.00027 | | | | |

| | ACC | | | | Disorder | | | | Helix | | | | Extended | | | | Loop | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | KW | C1 | C2 | Dp | KW | C1 | C2 | Dp | KW | C1 | C2 | Dp | KW | C1 | C2 | Dp | KW | C1 | C2 | Dp |
| | | 6 | 4 | 0.00012 | | | | | | | | | | 6 | 3 | 2.10E-10 | | | | |
| | | 6 | 5 | 1.00E-12 | | | | | | | | | | 6 | 4 | 2.10E-10 | | | | |
| | | | | | | | | | | | | | | 6 | 5 | < 2.00E-16 | | | | |
| **Cluster4** | 2.64E-13 | 4 | 1 | 4.80E-06 | 0.02282 | - | - | - | 0.0008189 | - | - | - | 1.60E-06 | 4 | 3 | 0.00038 | 5.99E-07 | 4 | 3 | 3.50E-07 |
| | | 3 | 2 | 0.00064 | | | | | | | | | | | | | | | | |
| | | 4 | 2 | 1.30E-06 | | | | | | | | | | | | | | | | |
| | | 6 | 3 | 2.30E-05 | | | | | | | | | | | | | | | | |
| | | 5 | 4 | 0.00066 | | | | | | | | | | | | | | | | |
| | | 6 | 4 | 6.50E-09 | | | | | | | | | | | | | | | | |

# Chapter 4 – Discussion

The past few decades a lot of research effort has been put on the identification and consequently filtering of regions with unusual amino acid composition. Upon the development of methods and tools for the systematic discovery of CBRs in the early 90s', the initial notion, was that they lack any biological significance mainly due to their unknown properties and troublesome statistical features (Altschul et al., 1994; S. Chavali et al., 2017; Wootton and Federhen, 1993). This built-up research of CBRs has led to a wealth of definitions and tools to be proposed, creating confusion on which tool or computational pipeline is best to follow when dealing with heavily biased genomes in comparative genomics analyses.

In this study, we proposed a suitable computational framework for handling CBRs in *computing extremely biased pan-genomes* and deploy the acquired knowledge for comparative genomics. Our approach enabled the *identification of an optimal masking strategy*, which is of critical importance, especially for heavily biased genomes, such as those of malaria parasites. Our data indicated that the robustness of comparative genomics computational pipeline is largely depending on the CBR filtering approach being followed.

A key feature of our proposed strategy is the employment of R-fraction (see equation **1**) prior clustering which, in simple words, is a normalized BLASTP bit score against the bit score of the ideal alignment of the query sequence by itself. The main idea behind R-fraction is that by comparing the query sequence by itself should score exceptionally high compared to the bit score obtained when compared against a large database such as NR. Hence, by normalizing BLAST bit-score against the self-alignment score we avoid any possible artifacts at the clustering step.

R-fraction proved instrumental in our analysis, as the best strategies consistently indicated that using R-fraction prior clustering could determine the most reliably identified protein families. For example, ranking of the discrete MCL modes using the Wilcoxon rank test between the worst performing filtering mode (query/database unmasked with no R-fraction prior clustering) and the constantly well performed masking modes (query/database masked with R-fraction prior clustering), we illustrated that masking modes with no R-fraction create statistically significant different protein families than the masking modes normalized by R-fraction prior clustering.

Specifically, we observed different protein families to be formed depending on which filtering mode we followed while at the same time, we identified 2498 protein families to cluster together independent of the masking mode (**Table 6**). We defined these protein families as robust and considered them as the most reliably identified protein families.

In our orthologous based clustering analysis, where we determined the robust clusters by combining clusters from our analysis and OrthoMCL, we observed 2275 (56.4%) core protein families describing, among others, Major *Plasmodium* protein families (MPFs) known for their medical importance in the development of an effective anti-malaria vaccine. Our results indicated that some of these MPFs are species specific families (e.g. PfEMP1) while others are restricted in certain *Plasmodium* lineages (e.g. phist and RAD proteins; **Table 9**). Earlier comparative genomics analyses using fewer *Plasmodium* species and slightly different methodology indicated different numbers of core protein families but similar evolutionary patterns for the MPFs (Cai et al., 2012, 2010; Carlton, 2006; Zilversmit et al., 2016). Our differences lie in the fact that we report only the robust core families (i.e. remained identical regardless of which filtering or clustering mode we followed). Even though, in some cases we observed families reported by Cartlon and colleagues (Carlton et al., 2008) to be absent from our analysis, the more comprehensive list of currently available genomes strengthens our results. One could consider this a disadvantage of our methodology since, failure to detect known MPFs suggests problematic comparative genomics pipeline. We argue that by reporting only the robust MPFs we provide the true structure of these protein families (i.e. a benchmark dataset) and thus, a more reliable dataset for structure, function and evolutionary studies.

Another key feature of the proposed comparative genomics analysis pipeline is the calculation of the Domain Composition Homogeneity (DCH; Figure 13) score where, we determine the homogeneity of protein domain architectures within each cluster based on how many different domains were found and the proteins having these domains. The significance of DCH score lies in its own definition which assess the different MCL modes from a biological point of view. Precisely, protein families with known functional regions and the identification of clusters with different combinations of domains can demonstrate the diverse range of proteins found in nature and therefore provide insights into their function (Finn et al., 2016). Hence, by calculating the DCH score we were able to determine the *Plasmodium* pan-genome protein domain architecture and validate our comparative genomics pipeline from a biological perspective. The protein domain

architecture analysis indicated that most clusters in each run are annotated by one or few Pfam domains and further affiliated that non-proper handling of CBRs creates artifacts in computational analyses.

We concluded that the optimal in silico strategy in comparative genomics is to mask both the query genome and database with CAST and apply the normalized BLAST bit scores (R-fraction) prior clustering.

Our view on the phylogenetic origin of *P. falciparum* has been constantly changing over the past few decades as more species are identified, sequenced and added to the tree and advances in molecular phylogenetic methods question older hypotheses. The earlier studies proposed an avian origin as a result of a relatively recent host switch based on Small Subunit (SSU) rRNA and circumsporozoite protein (CSP; (Escalante et al., 1998; Escalante and Ayala, 1994; Waters et al., 1991), while more recent studies have found evidence supporting that *P. falciparum* is closely related with the primate-infecting malaria parasites (Borner et al., 2016; Zheng et al., 2005; Zilversmit and Perkins, 2008). However, up to date, conclusive evidence indicate that the only species closely related to *P. falciparum* is *P. reichenowi* (Escalante and Ayala, 1994; Martinsen et al., 2008; Perkins and Schall, 2002), and that the two likely diverged from each other between 5 and 8 million years ago based on fossil dates of the human-chimpanzee split (Escalante and Ayala, 1994). Here, we constructed a Bayesian-inference tree of 21 *Plasmodium* species based on the *Plasmodium* core families in an effort to resolve this controversy. The robustness of our proposed phylogenetic tree lies in the methodology we followed. We used HMMER (Finn et al., 2015) to retrieve only the *Plasmodium* core clusters that composed of protein sequences that share statistically significant sequence similarity to *T. gondii* (our chosen outgroup species) proteins but also, mrBayes software implementing Bayesian inference across a wide range of phylogenetic and evolutionary models (Huelsenbeck and Ronquist, 2001). A comparison study between different phylogenetic approaches suggested that, the evolutionary pragmatism employed by the likelihood models is often compromised in order to improve the computational efficiency of the algorithms while the computationally intensive Bayesian approach, is complex, parameter rich, and is inferring phylogeny using a more realistic concept of evolution (Brooks et al., 2007). *Our results indicate the avian-origin hypothesis for P. falciparum* (see **Figure 18**) (Bensch et al., 2016; Pick et al., 2011; Tachibana et al., 2012) as the most probable one.

The Bayesian-inferenced *Plasmodium* species tree indicates that *P. gallinaceum* and *P. relictum* speciation predated the *P. falciparum* and *P. reichenowi* speciation. This observation is also supported by a recent Maximum Likelihood (ML) phylogeny of complete avian *Plasmodium* genomes (Böhme et al., 2018) . Thus, we suggest an avian origin for *P. falciparum* based on the core family-derived Plasmodial phylogenetic tree.

One of the central questions in evolutionary biology is the origin of new genes where several mechanisms (including gene duplication, retroposition, gene fusion/fission and horizontal gene transfer) have contributed to the birth of new genes (Kaessmann, 2010). All these proposed mechanisms require a preexisting gene that serves as the "mother" for the new gene but, several studies revealed de novo gene origin (i.e. a motherless mechanism) from non-coding DNA sequences in flies (McLysaght and Guerzoni, 2015; Reinhardt et al., 2013; Zhou et al., 2008), plants (Arendsee et al., 2014), bacteria (Cai et al., 2008; Fukuchi and Nishikawa, 2004; Fotis E. Psomopoulos et al., 2012) and parasites (Bozdech et al., 2008; Mukherjee et al., 2015; Sargeant et al., 2006). Such genes are true unique genes and may provide evidence on the molecular forces shaping each species evolution, pathogenicity and adaptation. In this particular set of experiments, we performed an in-depth comparative genomics analysis as an effort to explore and understand the molecular biology of the malaria parasites and at the same time, provide the necessary data and tools for devising more efficient techniques for novel discovery of drug/vaccine targets.

Controversial aspects of unique genes, e.g. shorter coding sequence, highly repetitive sequences, limited codon usage/amino acid composition and unstructured sequences, has led the scientific community arguing about their biological significance (Arendsee et al., 2014; Mukherjee et al., 2015; Verster et al., 2017; Wilson et al., 2005). We tested the hypothesis that unique genes show limited codon usage/amino acid composition by comparing the initial dataset of putative unique genes to the pan-genome (where we excluded the unique genes). By determining the differences between *Plasmodium* pan-genome and unique codon usage we could assess the evolutionary forces that shape the unique genes and devise more efficient techniques for structure and function determination. Our results indicated that *Plasmodium* putative unique proteins show a preference in simpler codon usage that reflects the highly A+T rich *Plasmodium* genomes and the driving forces of unique proteins' synthesis. Interestingly, malaria parasites *de*

*novo* genes show a preference in rare codons (i.e. codons that are not frequently used) and a higher percentage of the termination codons indicating that these unique genes are under specific evolutionary forces and mutation rates that ultimately will help us understand the genetic basis of each species phenotypic evolution and fitness adaptation. Despite the fascinating impact of these results we did not properly evaluate these observations, as we first choose to determine the genuine *Plasmodium* unique genes from those that are TRG, gene prediction artefacts or falsely set as unique due to failure of the gene prediction/annotation pipelines. However, an interesting follow-up project will be how much the codon usage varies between non-human and human-infecting *Plasmodium* species and how are these differences affecting each species pathogenicity. Additionally, by determining each species codon usage we could design more accurate algorithms and tools for determining which are the essential amino acids (that each species could not find in their respective hosts) for each *Plasmodium* species and thus, predicting the drug resistance of a genome.

Our initial dataset of the *Plasmodium* candidate unique proteins was composed of 1201 proteins comprising 1.11% of the *Plasmodium* pan-genome. Even though, advances in automatically gene prediction/annotation pipelines improved their prediction accuracy, it is still evident the effects of heavily biased genomes in their sensitivity and the need of manual curation of the newly sequenced genomes. Specifically, through our careful inspection of these "unique" protein sequences, we eliminated partial/fragmented sequences, detected gene prediction/annotation artifacts and exposed contaminated sequences.

Among the initial unique gene candidates, we identified 25 TRG genes illuminating interesting evolutionary patterns of the *Plasmodium* lineage specific traits. We must note, however, that these 25 TRGs are only a subset of all the *Plasmodium* TRGs. A future follow-up analysis is to determine the complete dataset of *Plasmodium* pan-genome TRGs, as studies have shown that taxonomically restricted genes are important for the evolution of lineage specific traits (Khalturin et al., 2009; Verster et al., 2017; Wilson et al., 2005). The main finding in this set of analysis was that the human-infecting parasites' proteins (all proteins from *P. ovale*) illustrate a wider phylogenetic profile as opposed to those from rodent-infecting malaria parasites. Based on the Presence/Absence heat-map tables we constructed (**Figure 34**), we observed that only two out of the six *P. ovale* proteins are restricted in human-infecting clade while, proteins from rodents and non-

human primate infecting parasites orthologous sequences were observed only in their respective clade. A recent study performed in *P. malariae* and *P. ovale* indicated that the host adaptations of these two species occurred over similar evolutionary timescales and that differences to the other malaria parasites regarding the gene content can be linked to their specific biology (Rutledge et al., 2017). Rutledge and colleagues suggested that the rodent-infecting malaria parasites could provide a closer model to the biology of *P. ovale* than other human-infecting species suggesting that there must have been an ancestral host switch from primates to rodents (Rutledge et al., 2017).

Another interesting aspect raised in recent studies is the fact that some *Plasmodium* species (e.g. the two *P. ovale* strains and *P. cynomolgi*) have been considered identical to other malaria parasites which, shows the limitation of morphology in species determination (Law, 2018; Rutledge et al., 2017; B. Singh et al., 2004). Scientists suspect that *P. cynomolgi* (the second malaria parasite, after *P. knowlesi*, found to switch hosts between human and non-human primates) infections have been occurring in people for years, but were misdiagnosed for another human malaria parasite, *P. vivax*, which looks similar (Law, 2018). Routinely-used malaria diagnosis tools may not be able to distinguish between the two species indicating an urgent need of specialized and more precise diagnostic tools (Law, 2018). Furthermore, almost all TRG proteins (as we noted in our results) are annotated either as “hypothetical protein” or as “conserved Plasmodium protein, unknown function” suggesting that improving both the gene prediction and functional annotation of the parasite genomes could be crucial in our efforts of rational design of highly effective drug/vaccine candidates and diagnostic tools.

The advent of genome sequencing technologies has opened exciting avenues for addressing long-standing questions in the evolution of cellular life (Dacks and Field, 2018; Harris et al., 2003; Promponas et al., 2016) and for developing ground-breaking biomedical, and biotechnological applications (Broder and Venter, 2000; Doble et al., 2017; Knight et al., 2012). Key, conserved biological processes are obvious targets for comparative analyses delineating the very deep evolutionary relations and for sketching a genomic outline of very early forms of life (Ouzounis and Kyrpides, 1996). Apparently, biological pathways and processes of central importance can be widely conserved among diverse life forms, hence their study is key towards unraveling deep phylogenies; N-glycosylation definitely fulfills the above criteria.

In the current study, we challenged the currently established notion that many of the OST complex subunits are not encoded in *Plasmodium* genomes. The initiation of this work was the unusually small Ost4p subunit, where the *Plasmodium* putative unique analysis detected it as *de novo* protein. However, systematic literature and database search provided strong evidence that all but one of the core OST complex subunits have homologs in *Plasmodium*. Microarray and EST/RNAseq data, *support the expression* of these newly characterized subunits in several malaria parasites. Moreover, the recent work reporting Ost3p/Ost6p homologs in *Plasmodium* spp. (Wang et al., 2017) independently supports our findings, even though this group presents evidence that in several developmental stages, the *P. berghei* homolog is translocated in the outer surfaces of the parasite. Thus, further work is necessary in order to elucidate the composition, stoichiometry and structure of the OST complex and also possible moonlighting functions of OST subunits in malaria parasites.

Our quest for identifying whether a Swp1/Ribophorin II homolog is encoded in *Plasmodium* genomes was unrewarding. Given the fact that these genes are essential in yeast and *Caenorhabditis elegans* (Kelleher and Gilmore, 2006) its absence from plasmodial species is puzzling, in the light of conservation of all other core subunits. The possibility that an evolutionary unrelated *Plasmodium* protein might compensate for the absence of this subunit could be investigated by exploiting the recently solved OST structures using fold recognition techniques which attempt to identify protein sequences compatible with a protein fold regardless of similarity at the sequence level (Casadio et al., 2007; Söding and Remmert, 2011) .

Our findings can be a starting point for studying the deep evolution of N-linked glycosylation in eukaryotes, with the aim of reconstructing the process in the last eukaryotic common ancestor. Even though previous efforts have been made towards this goal (e.g. (Lombard, 2016)), this challenging question is far from being resolved.

Moreover, we envisage that the recently determined structures of the yeast and mammalian OST complex and the updated view on the availability of OST subunits in *Plasmodium* (proposed in this work), will inspire new experimental and computational approaches for enhancing our understanding of the still puzzling issues regarding the importance of N-linked glycosylation in malaria parasites.

The A+T richness of malaria parasites genomes, induces an increased difficulty in genome sequencing projects and cloning in heterologous vector systems (Dunker et al., 2001; Malcolm J. Gardner et al., 2002; Promponas et al., 2000; Romero et al., 2000) which in turn, require special treatment for solving the 3D protein structure with trivial experimental procedures (Bannen et al., 2007; Coletta et al., 2010). Searching through PDB, we observed that only 1276 protein structures from *Plasmodium* species are currently deposited in PDB as opposed to 139,315 structures from other species (access date: 20/5/2018; (Berman et al., 2000)). The limited number of experimentally determined protein structures of the extremely biased *Plasmodium* species restrains our chances of correctly designating structural signatures of CBRs to protein sequences that will, ultimately, advance our chances on devising highly accurate and efficient *in-silico* structure prediction algorithms.

Accordingly, we selected a non-redundant dataset of experimentally determined 3D protein structures and using a reverse engineering approach we focused in designating structural signatures of CBRs to protein sequences. First, we determined both sequence and structural features of CBRs independently and then, using the k-means clustering algorithm we mapped the structural features back to the compositionally biased region. Hence, this approach enabled us to correctly assign specific secondary structure patterns to regions with composition bias, based on the over-represented residue type.

A contributing measure in our efforts to this reverse engineering approach is the calculations of Shannon entropy and local complexity of each CAST detected CBR-region. By calculating these two measures along with a normalized by region's length CAST score of each CBR we can determine the level of complexity of these regions and thus, properly map CBRs to a secondary structure pattern. For example, *we observed that certain CBRs would not be considered as low-complexity regions by SEG* as their region complexity and Shannon entropy is higher than the cutoff values set by SEG (using SEG default parameter set). Such residue types are Alanine, Glutamic acid, Serine and Threonine, indicating that CBRs rich in these amino acids tend to be more complex and have higher information content than other CBRs. Sub-clustering of the structural-derived clusters based solemnly on CBRs sequence features, strengthen the above observations as we observed statistically significant different sub-clusters to be formed. Specifically, our results portrayed both the structural preferences of CBRs (see **Figure 46**) but also, the sequence features these CBRs possess. Importantly, our findings can serve as a guideline for

optimizing disorder and structure prediction algorithms for CBR-containing proteins by incorporating different measures based on the causing bias residue type and region's sequence features. Furthermore, one can use the proposed methodology and refine his/her own dataset to evaluate the biological roles of CBRs based on specific structural or sequence signatures. For instance, we focused on Aspartic acid and Glutamic acid rich CBRs, which consistently are sub-clustered across the structure-derived clusters, discussing their possible biological roles. Although, its intriguing to determine and analyze *Plasmodium* structural and functional features of D-/E-rich CBRs specifically (due to the medical importance of Glutamic acid rich proteins to be used as biomarkers for the diagnosis of malaria parasites (Chou and Wang, 2015; Kattenberg et al., 2012)), we choose to present the general repertoire of biological functions/classifications of the experimentally solved structures deposited in PDB. As a future follow-up project, we plan an in-depth computational analysis both in *Plasmodium* species explicitly but also, of each cluster's detected CBRs focusing mainly on their biological significance.

Concerning the general behavior of CAST and our chosen threshold for CBR detection, we observed that *CBRs rich in hydrophobic residue types such as, Ile, Leu and Val, were significantly depleted*, a fact that might be explained by selection against the formation of aggregation-prone hyrdophobic patches. These residue types compose approximately 5.4%, 8.8% and 6.9% of the global composition (**Figure 42**). Even though, in BLOSUM62 Ile, Leu and Val have the same self-scoring value, CAST tend to mask the more abundant Leu residues. From this point of view these findings are expected, due to the fact that in a protein segment where Leu is in a larger fraction from Ile or Val, will give higher scoring and subsequently masked by CAST. An earlier work demonstrating the differences of each residue type between Simple Sequences and non-Simple sequences regarding their GC-content, suggested that Ile is one of the residue types found to be rich in organisms with low genomic GC-content (Subramanyam et al., 2006). On the contrary, Leu and Val found to display similar abundance patters in organisms with various GC-content. Taking a closer look at the standard genetic code, these findings are further supported by the AT-rich and GT-rich base composition of their codon. Regarding our chosen CAST threshold 25 instead of using the default cut-off value 40, our decision was based on an earlier study proposing CAST mode 25 as the optimal mode for detecting CBRs in PDB structures (Tamana et al., 2012).

Last, the fact that we treated missing structural residues as exposed could be considered as a disadvantage of our methodology. We treat those residues as exposed based mostly on the fact that it won't be any missing residues from the hydrophobic core because it should be packed in order to keep the whole protein together. An optimization of the script mapping protein sequence to structure is to consider the B-factors of residues flanking missing regions in order to correctly assign its exposure pattern. A recent study in missing structural residues demonstrated that flexible residues exhibiting high B-factors when exposed to the solvent, flanking the missing regions (Djinovic-Carugo and Carugo, 2015). Additionally, we could include structures solved by NMR spectroscopy or cryo-EM where it could help us distinguish missing residue-regions from genuine disorder regions since, both methods provide high resolution protein structures at the atomic level.

As a closing remark, the current thesis addressed numerous aspects of CBRs by employing experimentally solved structure datasets up to comparative genomics of the extremely biased malaria parasites genomes. We anticipate that our findings and proposed methodology provide novel knowledge to the research community that could further accelerate research both in the biomedical and biotechnological sectors. All software tools, results and data produced in this study are available to the research community, which we hope will (i) act as dissemination of our work, and (ii) inspire other research groups to expand this work.

## References


Aebi, M., 2013. N-linked protein glycosylation in the ER. Biochim. Biophys. Acta 1833, 2430–2437. https://doi.org/10.1016/j.bbamcr.2013.04.001

Akashi, H., Gojobori, T., 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc. Natl. Acad. Sci. 99, 3695–3700. https://doi.org/10.1073/pnas.062526999

Albà, M.M., Castresana, J., 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol. Biol. Evol. 22, 598–606. https://doi.org/10.1093/molbev/msi045

Albà, M.M., Laskowski, R.A., Hancock, J.M., 2002. Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinforma. Oxf. Engl. 18, 672–678.

Al-Khedery, B., Barnwell, J.W., Galinski, M.R., 1999. Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a P. knowlesi variant antigen. Mol. Cell 3, 131–141.

Altschul, S., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403–410.

Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C., 1994. Issues in searching molecular sequence databases. Nat. Genet. 6, 119–129. https://doi.org/10.1038/ng0294-119

Angiuoli, S.V., Dunning Hotopp, J.C., Salzberg, S.L., Tettelin, H., 2011. Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinformatics 12, 272. https://doi.org/10.1186/1471-2105-12-272

Ansari, H.R., Templeton, T.J., Subudhi, A.K., Ramaprasad, A., Tang, J., Lu, F., Naeem, R., Hashish, Y., Oguike, M.C., Benavente, E.D., Clark, T.G., Sutherland, C.J., Barnwell, J.W., Culleton, R., Cao, J., Pain, A., 2016. Genome-scale comparison of expanded gene families in Plasmodium ovale wallikeri and Plasmodium ovale curtisi with Plasmodium malariae and with other Plasmodium species. Int. J. Parasitol. 46, 685–696. https://doi.org/10.1016/j.ijpara.2016.05.009

Arendsee, Z.W., Li, L., Wurtele, E.S., 2014. Coming of age: orphan genes in plants. Trends Plant Sci. 19, 698–708. https://doi.org/10.1016/j.tplants.2014.07.003

Arisue, N., Hashimoto, T., Mitsui, H., Palacpac, N.M.Q., Kaneko, A., Kawai, S., Hasegawa, M., Tanabe, K., Horii, T., 2012. The Plasmodium apicoplast genome: conserved structure and close relationship of P. ovale to rodent malaria parasites. Mol. Biol. Evol. 29, 2095–2099. https://doi.org/10.1093/molbev/mss082

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. https://doi.org/10.1038/75556

Atkinson, H.J., Babbitt, P.C., 2009. An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. PLoS Comput. Biol. 5, e1000541. https://doi.org/10.1371/journal.pcbi.1000541

Auburn, S., Böhme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C.I., Berriman, M., Price, R.N., Otto, T.D., 2016. A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. 1, 4. https://doi.org/10.12688/wellcomeopenres.9876.1

Auer, M., Gremlich, H.U., Seifert, J.M., Daly, T.J., Parslow, T.G., Casari, G., Gstach, H., 1994. Helix-loop-helix motif in HIV-1 Rev. Biochemistry 33, 2988–2996.

Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O.S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J.C., Lawrence, C., Li, W., Pinney, D.F., Pulman, J.A., Roos, D.S., Shanmugasundram, A., Silva-Franco, F., Steinbiss, S., Stoeckert, C.J., Spruill, D., Wang, H., Warrenfeltz, S., Zheng, J., 2017. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. 45, D581–D591. https://doi.org/10.1093/nar/gkw1105

Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B.P., Caler, E.V., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Iodice, J., Kissinger, J.C., Kraemer, E.T., Li, W., Nayak, V., Pennington, C., Pinney, D.F., Pitts, B., Roos, D.S., Srinivasamoorthy, G., Stoeckert, C.J., Treatman, C., Wang, H., 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. Nucleic Acids Res. 39, D612-619. https://doi.org/10.1093/nar/gkq1006

Aurrecoechea, Cristina, Brestelli, J., Brunk, B.P., Carlton, J.M., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Morrison, H.G., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert, C.J., Sullivan, S., Treatman, C., Wang, H., 2009. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. Nucleic Acids Res. 37, D526-530. https://doi.org/10.1093/nar/gkn631

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Frank, I., John, I., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert Jr, C.J., Treatman, C., Wang, H., 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 37. https://doi.org/10.1093/nar/gkn814

Bai, L., Wang, T., Zhao, G., Kovach, A., Li, H., 2018. The atomic structure of a eukaryotic oligosaccharyltransferase complex. Nature 555, 328–333. https://doi.org/10.1038/nature25755

Banerjee, S., Vishwanath, P., Cui, J., Kelleher, D.J., Gilmore, R., Robbins, P.W., Samuelson, J., 2007. The evolution of N-glycan-dependent endoplasmic reticulum quality control factors for glycoprotein folding and degradation. Proc. Natl. Acad. Sci. U. S. A. 104, 11676–11681. https://doi.org/10.1073/pnas.0704862104

Bannen, R.M., Bingman, C.A., Phillips, G.N., 2007. Effect of low-complexity regions on protein structure determination. J. Struct. Funct. Genomics 8, 217–226. https://doi.org/10.1007/s10969-008-9039-6

Bapat, D., Huang, X., Gunalan, K., Preiser, P.R., 2011. Changes in Parasite Virulence Induced by the Disruption of a Single Member of the 235 kDa Rhoptry Protein Multigene Family of Plasmodium yoelii. PLOS ONE 6, e20170. https://doi.org/10.1371/journal.pone.0020170

Beck, T., 2006. The malaria parasite life cycle. [WWW Document]. URL https://www.ncbi.nlm.nih.gov/books/NBK5951/figure/malaria_LifeCycle/ (accessed 3.18.18).

Beeson, J.G., Drew, D.R., Boyle, M.J., Feng, G., Fowkes, F.J.I., Richards, J.S., 2016. Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. FEMS Microbiol. Rev. 40, 343–372. https://doi.org/10.1093/femsre/fuw001

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Bensch, S., Canbäck, B., DeBarry, J.D., Johansson, T., Hellgren, O., Kissinger, J.C., Palinauskas, V., Videvall, E., Valkiūnas, G., 2016. The Genome of Haemoproteus tartakovskyi and Its Relationship to Human Malaria Parasites. Genome Biol. Evol. 8, 1361–1373. https://doi.org/10.1093/gbe/evw081

Berer, K., Mues, M., Koutrolos, M., Rasbi, Z.A., Boziki, M., Johner, C., Wekerle, H., Krishnamoorthy, G., 2011. Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. Nature 479, 538–541. https://doi.org/10.1038/nature10554

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242.

Bhagavan, N.V., Ha, C.-E., 2011. Chapter 3 - Amino Acids, in: Essentials of Medical Biochemistry. Academic Press, San Diego, pp. 19–27. https://doi.org/10.1016/B978-0-12-095461-2.00003-5

Birkholtz, L.-M., Blatch, G., Coetzer, T.L., Hoppe, H.C., Human, E., Morris, E.J., Ngcete, Z., Oldfield, L., Roth, R., Shonhai, A., Stephens, L., Louw, A.I., 2008. Heterologous expression of plasmodial proteins for structural studies and functional annotation. Malar. J. 7, 197. https://doi.org/10.1186/1475-2875-7-197

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., Marceau, C., Janssen, H., Carette, J.E., Bennett, K.L., Colinge, J., Superti-Furga, G., Brummelkamp, T.R., 2015. Gene essentiality and synthetic lethality in haploid human cells. Science 350, 1092–1096. https://doi.org/10.1126/science.aac7557

Böhme, U., Otto, T.D., Cotton, J.A., Steinbiss, S., Sanders, M., Oyola, S.O., Nicot, A., Gandon, S., Patra, K.P., Herd, C., Bushell, E., Modrzynska, K.K., Billker, O., Vinetz, J.M., Rivero, A., Newbold, C.I., Berriman, M., 2018. Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution in birds and mammals. Genome Res. 28, 547–560. https://doi.org/10.1101/gr.218123.116

Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D., Kragelund, B.B., Best, R.B., Schuler, B., 2018. Extreme disorder in an ultrahigh-affinity protein complex. Nature 555, 61–66. https://doi.org/10.1038/nature25762

Borner, J., Pick, C., Thiede, J., Kolawole, O.M., Kingsley, M.T., Schulze, J., Cottontail, V.M., Wellinghausen, N., Schmidt-Chanasit, J., Bruchhaus, I., Burmester, T., 2016. Phylogeny of haemosporidian blood parasites revealed by a multi-gene approach. Mol. Phylogenet. Evol. 94, 221–231. https://doi.org/10.1016/j.ympev.2015.09.003

Bozdech, Z., Mok, S., Hu, G., Imwong, M., Jaidee, A., Russell, B., Ginsburg, H., Nosten, F., Day, N.P.J., White, N.J., Carlton, J.M., Preiser, P.R., 2008. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc. Natl. Acad. Sci. U. S. A. 105, 16290–16295. https://doi.org/10.1073/pnas.0807404105

Braunger, K., Pfeffer, S., Shrimal, S., Gilmore, R., Berninghausen, O., Mandon, E.C., Becker, T., Förster, F., Beckmann, R., 2018. Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. Science 360, 215–219. https://doi.org/10.1126/science.aar7899

Breitling, J., Aebi, M., 2013. N-Linked Protein Glycosylation in the Endoplasmic Reticulum. Cold Spring Harb. Perspect. Biol. 5. https://doi.org/10.1101/cshperspect.a013359

Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E., Karlin, S., 1992. Methods and algorithms for statistical analysis of protein sequences. Proc. Natl. Acad. Sci. U. S. A. 89, 2002–2006.

Broder, S., Venter, J.C., 2000. Whole genomes: the foundation of new biology and medicine. Curr. Opin. Biotechnol. 11, 581–585.

Brooks, D.R., Bilewitch, J., Charmaine, C., Evans, D.C., Folinsbee, K.E., Fröbisch, J., Halas, D., Hill, S., McLennan, D.A., Mattern, M., Tsuji, L.A., Ward, J.L., Wahlberg, N., Zamparo, D., Zanatta, D., 2007. Quantitative Phylogenetic Analysis in the 21st Century. Rev. Mex. Biodivers. 78, 225–252.

Brown, N.P., Leroy, C., Sander, C., 1998. MView: a web-compatible database search or multiple alignment viewer. Bioinforma. Oxf. Engl. 14, 380–381. https://doi.org/10.1093/bioinformatics/14.4.380

Bull, P.C., Abdi, A.I., 2016. The role of PfEMP1 as targets of naturally acquired immunity to childhood malaria: prospects for a vaccine. Parasitology 143, 171–186. https://doi.org/10.1017/S0031182015001274

Bushkin, G.G., Ratner, D.M., Cui, J., Banerjee, S., Duraisingh, M.T., Jennings, C.V., Dvorin, J.D., Gubbels, M.-J., Robertson, S.D., Steffen, M., O'Keefe, B.R., Robbins, P.W., Samuelson, J., 2010. Suggestive Evidence for Darwinian Selection against Asparagine-Linked Glycans of Plasmodium falciparum and Toxoplasma gondii. Eukaryot. Cell 9, 228–241. https://doi.org/10.1128/EC.00197-09

Cai, H., Gu, J., Wang, Y., 2010. Core genome components and lineage specific expansions in malaria parasites Plasmodium. BMC Genomics 11, S13. https://doi.org/10.1186/1471-2164-11-S3-S13

Cai, H., Zhou, Z., Gu, J., Wang, Y., 2012. Comparative Genomics and Systems Biology of Malaria ParasitesPlasmodium. Curr. Bioinforma. 7. https://doi.org/10.2174/157489312803900965

Cai, J., Zhao, R., Jiang, H., Wang, W., 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics 179, 487–496. https://doi.org/10.1534/genetics.107.084491

Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., Dunker, A.K., 2008. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. Protein Pept. Lett. 15, 956–963.

Cannarozzi, G., Cannarrozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., Barral, Y., 2010. A role for codon order in translation dynamics. Cell 141, 355–367. https://doi.org/10.1016/j.cell.2010.02.036

Carlton, J.M., 2006. Comparative genomics of Plasmodium species, in: Genomics and Evolution of Microbial Eukaryotes. Oxford Biology, pp. 33–47.

Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S.V., Merino, E.F., Amedeo, P., Cheng, Q., Coulson, R.M.R., Crabb, B.S., del Portillo, H.A., Essien, K., Feldblyum, T.V., Fernandez-Becerra, C., Gilson, P.R., Gueye, A.H., Guo, X., Kang'a, S., Kooij, T.W.A., Korsinczky, M., Meyer, E.V.-S., Nene, V., Paulsen, I., White, O., Ralph, S.A., Ren, Q., Sargeant, T.J., Salzberg, S.L., Stoeckert, C.J., Sullivan, S.A., Yamamoto, M.M., Hoffman, S.L., Wortman, J.R., Gardner, M.J., Galinski, M.R., Barnwell, J.W., Fraser-Liggett, C.M., 2008. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature 455, 757–763. https://doi.org/10.1038/nature07327

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway,

M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunningham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J., Carucci, D.J., 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 419, 512–519. https://doi.org/10.1038/nature01099

Casadio, R., Fariselli, P., Martelli, P.L., Tasco, G., 2007. Thinking the impossible: how to solve the protein folding problem with and without homologous structures and more. Methods Mol. Biol. Clifton NJ 350, 305–320.

Chang, S.-H., Chang, W.-L., Lu, C.-C., Tarn, W.-Y., 2014. Alanine repeats influence protein localization in splicing speckles and paraspeckles. Nucleic Acids Res. 42, 13788–13798. https://doi.org/10.1093/nar/gku1159

Chaudhari, N.M., Gupta, V.K., Dutta, C., 2016. BPGA- an ultra-fast pan-genome analysis pipeline. Sci. Rep. 6, 24373. https://doi.org/10.1038/srep24373

Chaudhry, S.R., Lwin, N., Phelan, D., Escalante, A.A., Battistuzzi, F.U., 2018. Comparative analysis of low complexity regions in Plasmodia. Sci. Rep. 8, 335. https://doi.org/10.1038/s41598-017-18695-y

Chavali, Sreenivas, Chavali, P.L., Chalancon, G., de Groot, N.S., Gemayel, R., Latysheva, N.S., Ing-Simmons, E., Verstrepen, K.J., Balaji, S., Babu, M.M., 2017. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat. Struct. Mol. Biol. 24, 765–777. https://doi.org/10.1038/nsmb.3441

Chavali, S., Chavali, P.L., Chalancon, G., De Groot, N.S., Gemayel, R., Latysheva, N.S., Ing-Simmons, E., Verstrepen, K.J., Balaji, S., Babu, M.M., 2017. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat. Struct. Mol. Biol. 24, 765–777. https://doi.org/10.1038/nsmb.3441

Chavan, M., Lennarz, W., 2006. The molecular basis of coupling of translocation and N-glycosylation. Trends Biochem. Sci. 31, 17–20. https://doi.org/10.1016/j.tibs.2005.11.010

Chen, F., Mackey, A.J., Stoeckert, C.J., Roos, D.S., 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34, D363–D368. https://doi.org/10.1093/nar/gkj123

Chi, J.H., Roos, J., Dean, N., 1996. The OST4 gene of Saccharomyces cerevisiae encodes an unusually small protein required for normal levels of oligosaccharyltransferase activity. J. Biol. Chem. 271, 3132–3140.

Chien, J.-T., Pakala, S.B., Geraldo, J.A., Lapp, S.A., Humphrey, J.C., Barnwell, J.W., Kissinger, J.C., Galinski, M.R., 2016. High-Quality Genome Assembly and Annotation for Plasmodium coatneyi, Generated Using Single-Molecule Real-Time PacBio Technology. Genome Announc. 4. https://doi.org/10.1128/genomeA.00883-16

Chou, C.-C., Wang, A.H.-J., 2015. Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry. Mol. Biosyst. 11, 2144–2151. https://doi.org/10.1039/C5MB00206K

Claverie, J.-M., States, D.J., 1993. Information enhancement methods for large scale sequence analysis. Comput. Chem. 17, 191–201. https://doi.org/10.1016/0097-8485(93)85010-A

Coletta, A., Pinney, J.W., Solís, D.Y.W., Marsh, J., Pettifer, S.R., Attwood, T.K., 2010. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst. Biol. 4, 43. https://doi.org/10.1186/1752-0509-4-43

Consortium, T.G.O., 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056. https://doi.org/10.1093/nar/gku1179

Cova, M., Rodrigues João, A., Smith, T.K., Izquierdo, L., 2015. Sugar activation and glycosylation in Plasmodium. Malar. J. 14.

Craig, A., Scherf, A., 2001. Molecules on the surface of the Plasmodium falciparum infected erythrocyte and their role in malaria pathogenesis and immune evasion. Mol. Biochem. Parasitol. 115, 129–143. https://doi.org/10.1016/S0166-6851(01)00275-4

Crick, S.L., Jayaraman, M., Frieden, C., Wetzel, R., Pappu, R.V., 2006. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. Proc. Natl. Acad. Sci. 103, 16764–16769. https://doi.org/10.1073/pnas.0608175103

Crosnier, C., Wanaguru, M., McDade, B., Osier, F.H., Marsh, K., Rayner, J.C., Wright, G.J., 2013. A Library of Functional Recombinant Cell-surface and Secreted P. falciparum Merozoite Proteins. Mol. Cell. Proteomics MCP 12, 3976–3986. https://doi.org/10.1074/mcp.O113.028357

Cuadrat, R.R.C., Da Serra Cruz, S.M., Tschoeke, D.A., Silva, E., Tosta, F., Jucá, H., Jardim, R., Campos, M.L.M., Mattoso, M., Dávila, A.M.R., 2014. An Orthology-Based Analysis of Pathogenic Protozoa Impacting Global Health: An Improved Comparative Genomics Approach with Prokaryotes and Model Eukaryote Orthologs. OMICS J. Integr. Biol. 18, 524–538. https://doi.org/10.1089/omi.2013.0172

Cunningham, D., Lawton, J., Jarra, W., Preiser, P., Langhorne, J., 2010. The pir multigene family of Plasmodium: antigenic variation and beyond. Mol. Biochem. Parasitol. 170, 65–73. https://doi.org/10.1016/j.molbiopara.2009.12.010

Dacks, J.B., Field, M.C., 2018. Evolutionary origins and specialisation of membrane transport. Curr. Opin. Cell Biol. 53, 70–76. https://doi.org/10.1016/j.ceb.2018.06.001

del Portillo, H.A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C.P., Schneider, N.K., Villalobos, J.M., Rajandream, M.-A., Harris, D., da Silva, L.H.P., Barrell, B., Lanzer, M., 2001. A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. Nature 410, 839–842. https://doi.org/10.1038/35071118

Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., Roucou, X., 2018. Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an mRNA. Proteomics 18, e1700058. https://doi.org/10.1002/pmic.201700058

Dietz, O., Rusch, S., Brand, F., Mundwiler-Pachlatko, E., Gaida, A., Voss, T., Beck, H.-P., 2014. Characterization of the Small Exported Plasmodium falciparum Membrane Protein SEMP1. PLOS ONE 9, e103272. https://doi.org/10.1371/journal.pone.0103272

Djinovic-Carugo, K., Carugo, O., 2015. Missing strings of residues in protein crystal structures. Intrinsically Disord. Proteins 3. https://doi.org/10.1080/21690707.2015.1095697

Doble, B., Schofield, D.J., Roscioli, T., Mattick, J.S., 2017. Prioritising the application of genomic medicine. NPJ Genomic Med. 2, 35. https://doi.org/10.1038/s41525-017-0037-0

Doerig, C., Rayner, J.C., Scherf, A., Tobin, A.B., 2015. Post-translational protein modifications in malaria parasites. Nat. Rev. Microbiol. 13, 160–172. https://doi.org/10.1038/nrmicro3402

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., Oggioni, M., Dunning Hotopp, J.C., Hu, F.Z., Riley, D.R., Covacci, A., Mitchell, T.J., Bentley, S.D., Kilian, M., Ehrlich, G.D., Rappuoli, R., Moxon, E.R., Masignani, V., 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol. 11, R107. https://doi.org/10.1186/gb-2010-11-10-r107

Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H., Spillane, C., 2011. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC Evol. Biol. 11, 47. https://doi.org/10.1186/1471-2148-11-47

Drummond, D.A., Wilke, C.O., 2009. The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. 10, 715–724. https://doi.org/10.1038/nrg2662

Dumax-Vorzet, A., Roboti, P., High, S., 2013. OST4 is a subunit of the mammalian oligosaccharyltransferase required for efficient N-glycosylation. J. Cell Sci. 126, 2595–2606. https://doi.org/10.1242/jcs.115410

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z., 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19, 26–59.

Dunn, O.J., 1964. Multiple Comparisons Using Rank Sums. Technometrics 6, 241–252. https://doi.org/10.1080/00401706.1964.10490181

Dunn, O.J., 1961. Multiple Comparisons among Means. J. Am. Stat. Assoc. 56, 52–64. https://doi.org/10.1080/01621459.1961.10482090

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.

Eberhardt, R.Y., Haft, D.H., Punta, M., Martin, M., O'Donovan, C., Bateman, A., 2012. AntiFam: a tool to help identify spurious ORFs in protein annotation. Database J. Biol. Databases Curation 2012, bas003. https://doi.org/10.1093/database/bas003

Eddy, S.R., 2011. Accelerated profile HMM searches. PLOS Comput Biol 7.

Elazar, A., Weinstein, J.J., Prilusky, J., Fleishman, S.J., 2016. Interplay between hydrophobicity and the positive-inside rule in determining membrane-protein topology. Proc. Natl. Acad. Sci. U. S. A. 113, 10340–10345. https://doi.org/10.1073/pnas.1605888113

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30, 1575–84.

Escalante, A.A., Ayala, F.J., 1994. Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proc. Natl. Acad. Sci. U. S. A. 91, 11373–11377.

Escalante, A.A., Freeland, D.E., Collins, W.E., Lal, A.A., 1998. The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. Proc. Natl. Acad. Sci. U. S. A. 95, 8124–8129.

Fairhurst, R.M., Nayyar, G.M.L., Breman, J.G., Hallett, R., Vennerstrom, J.L., Duong, S., Ringwald, P., Wellems, T.E., Plowe, C.V., Dondorp, A.M., 2012. Artemisinin-Resistant Malaria: Research Challenges, Opportunities, and Public Health Implications. Am. J. Trop. Med. Hyg. 87, 231–241. https://doi.org/10.4269/ajtmh.2012.12-0025

Farid, A., Malinovsky, F.G., Veit, C., Schoberer, J., Zipfel, C., Strasser, R., 2013. Specialized roles of the conserved subunit OST3/6 of the oligosaccharyltransferase complex in

innate immunity and tolerance to abiotic stresses. Plant Physiol. 162, 24–38. https://doi.org/10.1104/pp.113.215509
Fasta to Nexus Sequence Converter [WWW Document], n.d. URL http://sequenceconversion.bugaco.com/converter/biology/sequences/fasta_to_nexus.php (accessed 10.23.17).
Favaloro, J.M., Culvenor, J.G., Anders, R.F., Kemp, D.J., 1993. A Plasmodium chabaudi antigen located in the parasitophorous vacuole membrane. Mol. Biochem. Parasitol. 62, 263–270.
Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., Eddy, S.R., 2015. HMMER web server: 2015 update. Nucleic Acids Res. 43, W30–W38. https://doi.org/10.1093/nar/gkv397
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285. https://doi.org/10.1093/nar/gkv1344
Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L., Bateman, A., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247–D251. https://doi.org/10.1093/nar/gkj149
Fisher, R.A., 1992. Statistical Methods for Research Workers, in: Kotz, S., Johnson, N.L. (Eds.), Breakthroughs in Statistics: Methodology and Distribution, Springer Series in Statistics. Springer New York, New York, NY, pp. 66–70. https://doi.org/10.1007/978-1-4612-4380-9_6
Fisher, R.A., 1922. On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. J. R. Stat. Soc. 85, 87–94. https://doi.org/10.2307/2340521
Florent, I., Maréchal, E., Gascuel, O., Bréhélin, L., 2010. Bioinformatic strategies to provide functional clues to the unknown genes in Plasmodium falciparum genome. Parasite 17, 273–283. https://doi.org/10.1051/parasite/2010174273
Fornace, K.M., Abidin, T.R., Alexander, N., Brock, P., Grigg, M.J., Murphy, A., William, T., Menon, J., Drakeley, C.J., Cox, J., 2016. Association between Landscape Factors and Spatial Patterns of Plasmodium knowlesi Infections in Sabah, Malaysia. Emerg. Infect. Dis. 22, 201–208. https://doi.org/10.3201/eid2202.150656
Fougère, A., Jackson, A.P., Bechtsi, D.P., Braks, J.A.M., Annoura, T., Fonager, J., Spaccapelo, R., Ramesar, J., Chevalley-Maurel, S., Klop, O., van der Laan, A.M.A., Tanke, H.J., Kocken, C.H.M., Pasini, E.M., Khan, S.M., Böhme, U., Van Ooij, C., Otto, T.D., Janse, C.J., Franke-Fayard, B., 2017. Correction: Variant Exported Blood-Stage Proteins Encoded by Plasmodium Multigene Families Are Expressed in Liver Stages Where They Are Exported into the Parasitophorous Vacuole. PLoS Pathog. 13, e1006128. https://doi.org/10.1371/journal.ppat.1006128
Frech, C., Chen, N., 2013. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. BMC Genomics 14, 427. https://doi.org/10.1186/1471-2164-14-427
Frech, C., Chen, N., 2011. Genome Comparison of Human and Non-Human Malaria Parasites Reveals Species Subset-Specific Genes Potentially Linked to Human Disease. PLOS Comput Biol 7, e1002320. https://doi.org/10.1371/journal.pcbi.1002320
Fredrick, K., Ibba, M., 2010. How the sequence of a gene can tune its translation. Cell 141, 227–229. https://doi.org/10.1016/j.cell.2010.03.033

Frugier, M., Bour, T., Ayach, M., Santos, M.A.S., Rudinger-Thirion, J., Théobald-Dietrich, A., Pizzi, E., 2010. Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. FEBS Lett. 584, 448–454. https://doi.org/10.1016/j.febslet.2009.11.004

Fukuchi, S., Nishikawa, K., 2004. Estimation of the number of authentic orphan genes in bacterial genomes. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes 11, 219–231, 311–313.

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J.C., Mackey, A.J., Pinney, D.F., Roos, D.S., Stoeckert, C.J., Wang, H., Brunk, B.P., 2008. ToxoDB: an integrated Toxoplasma gondii database resource. Nucleic Acids Res. 36, D553-556. https://doi.org/10.1093/nar/gkm981

Gardner, Malcolm J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M.A., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419, 498–511. https://doi.org/10.1038/nature01097

Gardner, M. J., Shallom, S.J., Carlton, J.M., Salzberg, S.L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L.M., Smith, H.O., Adams, M.D., Venter, J.C., Carucci, D.J., Hoffman, S.L., Fraser, C.M., 2002. Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature 419, 531–534. https://doi.org/10.1038/nature01094

Gayen, S., Kang, C., 2011. Solution structure of a human minimembrane protein Ost4, a subunit of the oligosaccharyltransferase complex. Biochem. Biophys. Res. Commun. 409, 572–576. https://doi.org/10.1016/j.bbrc.2011.05.050

Genetics Society of America, 2018. Pseudogenes. Encyclopedia.com, Genetics.

Genome, National Center for Biotechnology Information, 2018. Genomes - Genome - NCBI [WWW Document]. URL https://www.ncbi.nlm.nih.gov/genome/genomes/33 (accessed 5.18.17).

Golding, G.B., 1999. Simple sequence is abundant in eukaryotic proteins. Protein Sci. Publ. Protein Soc. 8, 1358–1361. https://doi.org/10.1110/ps.8.6.1358

Gotliv, B.-A., Kessler, N., Sumerel, J.L., Morse, D.E., Tuross, N., Addadi, L., Weiner, S., 2005. Asprich: A novel aspartic acid-rich protein family from the prismatic shell matrix of the bivalve Atrina rigida. Chembiochem Eur. J. Chem. Biol. 6, 304–314. https://doi.org/10.1002/cbic.200400221

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

Graham, D.B., Lefkovith, A., Deelen, P., de Klein, N., Varma, M., Boroughs, A., Desch, A.N., Ng, A.C.Y., Guzman, G., Schenone, M., Petersen, C.P., Bhan, A.K., Rivas, M.A., Daly, M.J., Carr, S.A., Wijmenga, C., Xavier, R.J., 2016. TMEM258 Is a Component of the Oligosaccharyltransferase Complex Controlling ER Stress and Intestinal Inflammation. Cell Rep. 17, 2955–2965. https://doi.org/10.1016/j.celrep.2016.11.042

Grantham, R., Gautier, C., Gouy, M., 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 8, 1893–1912.

Grignaschi, E., Cereghetti, G., Grigolato, F., Kopp, M.R.G., Caimi, S., Faltova, L., Saad, S., Peter, M., Arosio, P., 2018. A hydrophobic low-complexity region regulates aggregation of the yeast pyruvate kinase Cdc19 into amyloid-like aggregates *in vitro*. J. Biol. Chem. 293, 11424–11432. https://doi.org/10.1074/jbc.RA117.001628

Guha-Niyogi, A., Sullivan, D.R., Turco, S.J., 2001. Glycoconjugate structures of parasitic protozoa. Glycobiology 11, 45R-59R.

Gunalan, K., Gao, X., Yap, S.S.L., Huang, X., Preiser, P.R., 2013. The role of the reticulocyte-binding-like protein homologues of Plasmodium in erythrocyte sensing and invasion. Cell. Microbiol. 15, 35–44. https://doi.org/10.1111/cmi.12038

Gundersen, V., 2010. Protein aggregation in Parkinson's disease. Acta Neurol. Scand. Suppl. 82–87. https://doi.org/10.1111/j.1600-0404.2010.01382.x

Gupta, R., Jung, E., Brunak, S., 2004. Prediction of N-glycosylation sites in human proteins. Prep.

Hagner, S.C., Misof, B., Maier, W.A., Kampen, H., 2007. Bayesian analysis of new and old malaria parasite DNA sequence data demonstrates the need for more phylogenetic signal to clarify the descent of Plasmodium falciparum. Parasitol. Res. 101, 493–503. https://doi.org/10.1007/s00436-007-0499-6

Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W.A., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M.A., Ormond, D., Doggett, J., Trueman, H.E., Mendoza, J., Bidwell, S.L., Rajandream, M.-A., Carucci, D.J., Yates, J.R., Kafatos, F.C., Janse, C.J., Barrell, B., Turner, C.M.R., Waters, A.P., Sinden, R.E., 2005. A Comprehensive Survey of the Plasmodium Life Cycle by Genomic, Transcriptomic, and Proteomic Analyses. Science 307, 82–86. https://doi.org/10.1126/science.1103717

Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S., Barron, A., Brooks, K., Buckee, C.O., Burrows, C., Cherevach, I., Chillingworth, C., Chillingworth, T., Christodoulou, Z., Clark, L., Clark, R., Corton, C., Cronin, A., Davies, R., Davis, P., Dear, P., Dearden, F., Doggett, J., Feltwell, T., Goble, A., Goodhead, I., Gwilliam, R., Hamlin, N., Hance, Z., Harper, D., Hauser, H., Hornsby, T., Holroyd, S., Horrocks, P., Humphray, S., Jagels, K., James, K.D., Johnson, D., Kerhornou, A., Knights, A., Konfortov, B., Kyes, S., Larke, N., Lawson, D., Lennard, N., Line, A., Maddison, M., McLean, J., Mooney, P., Moule, S., Murphy, L., Oliver, K., Ormond, D., Price, C., Quail, M.A., Rabbinowitsch, E., Rajandream, M.-A., Rutter, S., Rutherford, K.M., Sanders, M., Simmonds, M., Seeger, K., Sharp, S., Smith, R., Squares, R., Squares, S., Stevens, K., Taylor, K., Tivey, A., Unwin, L., Whitehead, S., Woodward, J., Sulston, J.E., Craig, A., Newbold, C., Barrell, B.G., 2002. Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. Nature 419, 527–531. https://doi.org/10.1038/nature01095

Han, M.V., Zmasek, C.M., 2009. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics 10, 356. https://doi.org/10.1186/1471-2105-10-356

Hancock, J.M., Armstrong, J.S., 1994. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput. Appl. Biosci. CABIOS 10, 67–70.

Harbi, D., Kumar, M., Harrison, P.M., 2011. LPS-annotate: complete annotation of compositionally biased regions in the protein knowledgebase. Database J. Biol. Databases Curation 2011. https://doi.org/10.1093/database/baq031
Harris, J.K., Kelley, S.T., Spiegelman, G.B., Pace, N.R., 2003. The genetic core of the universal ancestor. Genome Res. 13, 407–412. https://doi.org/10.1101/gr.652803
Harrison, P.M., 2017. fLPS: Fast discovery of compositional biases for the protein universe. BMC Bioinformatics 18, 476. https://doi.org/10.1186/s12859-017-1906-3
Hayakawa, T., Culleton, R., Otani, H., Horii, T., Tanabe, K., 2008. Big bang in the evolution of extant malaria parasites. Mol. Biol. Evol. 25, 2233–2239. https://doi.org/10.1093/molbev/msn171
He, D., Parkinson, J., 2008. SubSeqer: a graph-based approach for the detection and identification of repetitive elements in low-complexity sequences. Bioinforma. Oxf. Engl. 24, 1016–1017. https://doi.org/10.1093/bioinformatics/btn073
Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C., Gao, X., Kaluskar, N., Rhodes, P., Wang, S., He, C.-Z., Su, Y., Miller, J., Kraemer, E., Kissinger, J.C., 2006. CryptoDB: a Cryptosporidium bioinformatics resource update. Nucleic Acids Res. 34, D419-422. https://doi.org/10.1093/nar/gkj078
Heizer, E.M., Raiford, D.W., Raymer, M.L., Doom, T.E., Miller, R.V., Krane, D.E., 2006. Amino Acid Cost and Codon-Usage Biases in 6 Prokaryotic Genomes: A Whole-Genome Analysis. Mol. Biol. Evol. 23, 1670–1680. https://doi.org/10.1093/molbev/msl029
Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287–299. https://doi.org/10.1146/annurev.genet.42.110807.091442
Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., von Heijne, G., 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 433, 377–381. https://doi.org/10.1038/nature03216
HMMER [WWW Document], n.d. URL http://hmmer.org/ (accessed 10.23.17).
Hollenstein, K., Dawson, R.J.P., Locher, K.P., 2007. Structure and mechanism of ABC transporter proteins. Curr. Opin. Struct. Biol. 17, 412–418. https://doi.org/10.1016/j.sbi.2007.07.003
Höps, W., Jeffryes, M., Bateman, A., 2018. Gene Unprediction with Spurio: A tool to identify spurious protein sequences. F1000Research 7, 261. https://doi.org/10.12688/f1000research.14050.1
Horvath, M.P., 2013. Evolution of Telomere Binding Proteins. Landes Bioscience.
Hubbard, S.J., Thornton, J.M., 1993. NACCESS, Computer Program. Department of Biochemistry and Molecular Biology, University College London.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218. https://doi.org/10.1007/BF01908075
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinforma. Oxf. Engl. 17, 754–755.
Hughes, A.L., 2004. The Evolution of Amino Acid Repeat Arrays in Plasmodium and Other Organisms. J. Mol. Evol. 59, 528–535. https://doi.org/10.1007/s00239-004-2645-4
Hung, L.W., Wang, I.X., Nikaido, K., Liu, P.Q., Ames, G.F., Kim, S.H., 1998. Crystal structure of the ATP-binding subunit of an ABC transporter. Nature 396, 703–707. https://doi.org/10.1038/25393
Huntley, M.A., Golding, G.B., 2002. Simple sequences are rare in the Protein Data Bank. Proteins 48, 134–140. https://doi.org/10.1002/prot.10150

Imam, M., Singh, S., Kaushik, N.K., Chauhan, V.S., 2014. Plasmodium falciparum Merozoite Surface Protein 3. J. Biol. Chem. 289, 3856–3868. https://doi.org/10.1074/jbc.M113.520239

Iwagami, M., Hwang, S.-Y., Fukumoto, M., Hayakawa, T., Tanabe, K., Kim, S.-H., Kho, W.-G., Kano, S., 2010. Geographical origin of Plasmodium vivax in the Republic of Korea: haplotype network analysis based on the parasite's mitochondrial genome. Malar. J. 9, 184. https://doi.org/10.1186/1475-2875-9-184

Janssen, C.S., Barrett, M.P., Turner, C.M.R., Phillips, R.S., 2002. A large gene family for putative variant antigens shared by human and rodent malaria parasites. Proc. R. Soc. B Biol. Sci. 269, 431–436. https://doi.org/10.1098/rspb.2001.1903

Janssen, C.S., Phillips, R.S., Turner, C.M.R., Barrett, M.P., 2004. Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. Nucleic Acids Res. 32, 5712–5720. https://doi.org/10.1093/nar/gkh907

Janssen, P., Enright, A.J., Audit, B., Cases, I., Goldovsky, L., Harte, N., Kunin, V., Ouzounis, C.A., 2003. COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics. Bioinformatics 19, 1451–1452. https://doi.org/10.1093/bioinformatics/btg161

Jongwutiwes, S., Putaporntip, C., Iwasaki, T., Ferreira, M.U., Kanbara, H., Hughes, A.L., 2005. Mitochondrial genome sequences support ancient population expansion in Plasmodium vivax. Mol. Biol. Evol. 22, 1733–1739. https://doi.org/10.1093/molbev/msi168

Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C., Vriend, G., 2011. A series of PDB related databases for everyday needs. Nucleic Acids Res. 39, D411–D419. https://doi.org/10.1093/nar/gkq1105

Jorda, J., Kajava, A.V., 2010. Protein homorepeats sequences, structures, evolution, and functions. Adv. Protein Chem. Struct. Biol. 79, 59–88. https://doi.org/10.1016/S1876-1623(10)79002-7

Kabsch, W.G., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kaessmann, H., 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20, 1313–1326. https://doi.org/10.1101/gr.101386.109

Kaneko, I., Iwanaga, S., Kato, T., Kobayashi, I., Yuda, M., 2015. Genome-Wide Identification of the Target Genes of AP2-O, a Plasmodium AP2-Family Transcription Factor. PLOS Pathog. 11, e1004905. https://doi.org/10.1371/journal.ppat.1004905

Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U. S. A. 87, 2264–2268.

Karlin, S., Brocchieri, L., Bergman, A., Mrázek, J., Gentles, A.J., 2002. Amino acid runs in eukaryotic proteomes and disease associations. Proc. Natl. Acad. Sci. 99, 333–338. https://doi.org/10.1073/pnas.012608599

Katsani, K.R., Irimia, M., Karapiperis, C., Scouras, Z.G., Blencowe, B.J., Promponas, V.J., Ouzounis, C.A., 2014. Functional genomics evidence unearths new moonlighting roles of outer ring coat nucleoporins. Sci. Rep. 4, 4655. https://doi.org/10.1038/srep04655

Kattenberg, J.H., Versteeg, I., Migchelsen, S.J., González, I.J., Perkins, M.D., Mens, P.F., Schallig, H.D., 2012. New developments in malaria diagnostics. mAbs 4, 120–126. https://doi.org/10.4161/mabs.4.1.18529

Kavishe, R.A., van den Heuvel, J.M., van de Vegte-Bolmer, M., Luty, A.J., Russel, F.G., Koenderink, J.B., 2009. Localization of the ATP-binding cassette (ABC) transport proteins PfMRP1, PfMRP2, and PfMDR5 at the Plasmodium falciparum plasma membrane. Malar. J. 8, 205. https://doi.org/10.1186/1475-2875-8-205

Kay, B.K., Williamson, M.P., Sudol, M., 2000. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 14, 231–241.

Kelleher, D.J., Gilmore, R., 2006. An evolving view of the eukaryotic oligosaccharyltransferase. Glycobiology 16, 47R-62R. https://doi.org/10.1093/glycob/cwj066

Kelleher, D.J., Gilmore, R., 1994. The Saccharomyces cerevisiae oligosaccharyltransferase is a protein complex composed of Wbp1p, Swp1p, and four additional polypeptides. J. Biol. Chem. 269, 12908–12917.

Kessler, M.M., Zeng, Q., Hogan, S., Cook, R., Morales, A.J., Cottarel, G., 2003. Systematic discovery of new genes in the Saccharomyces cerevisiae genome. Genome Res. 13, 264–271. https://doi.org/10.1101/gr.232903

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., Bosch, T.C.G., 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25, 404–413. https://doi.org/10.1016/j.tig.2009.07.006

Kimura, E.A., Couto, A.S., Peres, V.J., Casal, O.L., Katzin, A.M., 1996. N-Linked Glycoproteins Are Related to Schizogony of the Intraerythrocytic Stage in Plasmodium falciparum. J. Biol. Chem. 271, 14452–14461. https://doi.org/10.1074/jbc.271.24.14452

Kirmitzoglou, I., 2014. Development of algorithms and software for unravelling the biological role of low complexity regions in protein sequences (PhD Thesis). University of Cyprus, Nicosia, Cyprus.

Kirmitzoglou, I., Promponas, V.J., 2015. LCR-eXXXplorer: a web platform to search, visualize and share data for low complexity regions in protein sequences. Bioinformatics 31, 2208–2210. https://doi.org/10.1093/bioinformatics/btv115

Kissinger, J.C., DeBarry, J., 2011. Genome cartography: charting the apicomplexan genome. Trends Parasitol. 27, 345–354. https://doi.org/10.1016/j.pt.2011.03.006

Klein, E.Y., 2013. Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread. Int. J. Antimicrob. Agents 41, 311–317. https://doi.org/10.1016/j.ijantimicag.2012.12.007

Knauer, R., Lehle, L., 1999. The oligosaccharyltransferase complex from yeast. Biochim. Biophys. Acta 1426, 259–273.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., Bailey, M.J., Gordon, J.I., Kowalchuk, G.A., Gilbert, J.A., 2012. Unlocking the potential of metagenomics through replicated experimental design. Nat. Biotechnol. 30, 513–520. https://doi.org/10.1038/nbt.2235

Konopka, A.K., Owens, J., 1990. Complexity charts can be used to map functional domains in DNA. Genet. Anal. Tech. Appl. 7, 35–38.

Kooij, T.W.A., Carlton, J.M., Bidwell, S.L., Hall, N., Ramesar, J., Janse, C.J., Waters, A.P., 2005. A Plasmodium whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. PLoS Pathog. 1, e44. https://doi.org/10.1371/journal.ppat.0010044

Koonin, E.V., Mushegian, A.R., 1996. Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. Curr. Opin. Genet. Dev. 6, 757–762.

Kreil, D.P., Ouzounis, C.A., 2003. Comparison of sequence masking algorithms and the detection of biased protein sequence regions. Bioinformatics 19, 1672–1681. https://doi.org/10.1093/bioinformatics/btg212

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kruskal, W.H., Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. J. Am. Stat. Assoc. 47, 583–621. https://doi.org/10.1080/01621459.1952.10483441

Krzywinski, M., Altman, N., 2014. Points of Significance: Visualizing samples with box plots [WWW Document]. Nat. Methods. https://doi.org/10.1038/nmeth.2813

Kudla, G., Murray, A.W., Tollervey, D., Plotkin, J.B., 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324, 255–258. https://doi.org/10.1126/science.1170160

Kumar, A.S., Sowpati, D.T., Mishra, R.K., 2016. Single Amino Acid Repeats in the Proteome World: Structural, Functional, and Evolutionary Insights. PLOS ONE 11, e0166854. https://doi.org/10.1371/journal.pone.0166854

Kumar, M., Kumari, B., Kumar, R., 2017. Structural Analysis of Low Complexity Regions of Proteins. Can J Biotech Volume 1, 219.

Kumar, P., Bansal, M., 2012. HELANAL-Plus: a web server for analysis of helix geometry in protein structures. J. Biomol. Struct. Dyn. 30, 773–783.

Kumari, B., Kumar, R., Kumar, M., 2015. Low complexity and disordered regions of proteins have different structural and amino acid preferences. Mol. Biosyst. 11, 585–594. https://doi.org/10.1039/c4mb00425f

Kuznetsov, I.B., Hwang, S., 2006. A novel sensitive method for the detection of user-defined compositional bias in biological sequences. Bioinforma. Oxf. Engl. 22, 1055–1063. https://doi.org/10.1093/bioinformatics/btl049

Law, Y.-H., 2018. Rare human outbreak of monkey malaria detected in Malaysia [WWW Document]. Nature. https://doi.org/10.1038/d41586-018-04121-4

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55, 379-IN4. https://doi.org/10.1016/0022-2836(71)90324-X

Lee, K.-S., Divis, P.C.S., Zakaria, S.K., Matusop, A., Julin, R.A., Conway, D.J., Cox-Singh, J., Singh, B., 2011. Plasmodium knowlesi: Reservoir Hosts and Tracking the Emergence in Humans and Macaques. PLOS Pathog. 7, e1002015. https://doi.org/10.1371/journal.ppat.1002015

Li, Hua, Chavan, M., Schindelin, H., Lennarz, W.J., Li, Huilin, 2008. Structure of the oligosaccharyl transferase complex at 12 A resolution. Struct. Lond. Engl. 1993 16, 432–440. https://doi.org/10.1016/j.str.2007.12.013

Li, L., Foster, C.M., Gan, Q., Nettleton, D., James, M.G., Myers, A.M., Wurtele, E.S., 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. Plant J. 58, 485–498. https://doi.org/10.1111/j.1365-313X.2009.03793.x

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13, 2178–2189. https://doi.org/10.1101/gr.1224503

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y.M., Buso, N., Lopez, R., 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 43, W580–W584. https://doi.org/10.1093/nar/gkv279

Li, X., Kahveci, T., 2006. A Novel algorithm for identifying low-complexity regions in a protein sequence. Bioinforma. Oxf. Engl. 22, 2980–2987. https://doi.org/10.1093/bioinformatics/btl495

Li, Y., Fanning, A.S., Anderson, J.M., Lavie, A., 2005. Structure of the conserved cytoplasmic C-terminal domain of occludin: identification of the ZO-1 binding surface. J. Mol. Biol. 352, 151–164. https://doi.org/10.1016/j.jmb.2005.07.017

Liu, J., Istvan, E.S., Gluzman, I.Y., Gross, J., Goldberg, D.E., 2006. Plasmodium falciparum ensures its amino acid supply with multiple acquisition pathways and redundant proteolytic enzyme systems. Proc. Natl. Acad. Sci. 103, 8840–8845. https://doi.org/10.1073/pnas.0601876103

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., 2000. Hierarchical Structure of Proteins. Mol. Cell Biol. 4th Ed.

Lombard, J., 2016. The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. Biol. Direct 11, 36. https://doi.org/10.1186/s13062-016-0137-2

Lozupone, C.A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J.K., Gordon, J.I., Knight, R., 2013. Meta-analyses of studies of the human microbiota. Genome Res. 23, 1704–1714. https://doi.org/10.1101/gr.151803.112

Ludin, P., Woodcroft, B., Ralph, S.A., Mäser, P., 2012. In silico prediction of antimalarial drug target candidates. Int. J. Parasitol. Drugs Drug Resist. 2, 191–199. https://doi.org/10.1016/j.ijpddr.2012.07.002

Macedo, C.S., Schwarz, R.T., Todeschini, A.R., Previato, J.O., Mendonça-Previato, L., 2010. Overlooked post-translational modifications of proteins in Plasmodium falciparum: N- and O-glycosylation - A Review. Mem. Inst. Oswaldo Cruz 105, 949–956. https://doi.org/10.1590/S0074-02762010000800001

Makino, S., Qu, J.N., Uemori, K., Ichikawa, H., Ogura, T., Matsuzawa, H., 1997. A silent mutation in the ftsH gene of Escherichia coli that affects FtsH protein production and colicin tolerance. Mol. Gen. Genet. MGG 254, 578–583.

Marin, M., 2008. Folding at the rhythm of the rare codon beat. Biotechnol. J. 3, 1047–1057. https://doi.org/10.1002/biot.200800089

Marín-Menéndez, A., Monaghan, P., Bell, A., 2012. A family of cyclophilin-like molecular chaperones in Plasmodium falciparum. Mol. Biochem. Parasitol. 184, 44–47. https://doi.org/10.1016/j.molbiopara.2012.04.006

Martinsen, E.S., Perkins, S.L., Schall, J.J., 2008. A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): evolution of life-history traits and host switches. Mol. Phylogenet. Evol. 47, 261–273. https://doi.org/10.1016/j.ympev.2007.11.012

McColl, D.J., Silva, A., Foley, M., Kun, J.F., Favaloro, J.M., Thompson, J.K., Marshall, V.M., Coppel, R.L., Kemp, D.J., Anders, R.F., 1994. Molecular variation in a novel polymorphic antigen associated with Plasmodium falciparum merozoites. Mol. Biochem. Parasitol. 68, 53–67.

McCutchan, T.F., Kissinger, J.C., Touray, M.G., Rogers, M.J., Li, J., Sullivan, M., Braga, E.M., Krettli, A.U., Miller, L.H., 1996. Comparison of circumsporozoite proteins from avian and mammalian malarias: biological and phylogenetic implications. Proc. Natl. Acad. Sci. 93, 11889–11894.

McGill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of Box Plots. Am. Stat. 32, 12–16. https://doi.org/10.2307/2683468

McLysaght, A., Guerzoni, D., 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Phil Trans R Soc B 370, 20140332. https://doi.org/10.1098/rstb.2014.0332

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., Lopez, R., 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 41, W597-600. https://doi.org/10.1093/nar/gkt376

Michelitsch, M.D., Weissman, J.S., 2000. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. Proc. Natl. Acad. Sci. U. S. A. 97, 11910–11915. https://doi.org/10.1073/pnas.97.22.11910

Mier, P., Paladin, L., Tamana, S., Petrosian, S., Hajdu-Soltesz, B., Urbanek, A., Gruca, A., Plewczynski, D., Grynberg, M., Bernado, P., Gaspari, Z., Ouzounis, C., Promponas, V.J., Kajava, A.V., Hancock, J.M., Tosatto, S., Dosztanyi, Z., Andrade-Navarro, M.A., 2018. Disentangling the complexity of low complexity proteins Pablo Mier1, Lisanna Paladin2, Stella Tamana3, Sophia Petrosian4, Borbála Hajdu-Soltész5, Annika Urbanek6, Aleksandra Gruca7, Dariusz Plewczynski8, Marcin Grynberg9, Pau Bernadó6, Zoltán Gáspári10, Christos Ouzounis4, Vasilis J. Promponas3, Andrey V. Kajava11,12, John M. Hancock13,14, Silvio Tosatto2, Zsuzsanna Dosztanyi5, Miguel A. Andrade-Navarro1. Brief. Bioinform. (In press).

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proc. Gatew. Comput. Environ. Workshop GCE New Orleans, LA, 1–8.

Mohd Yusuf, S.N.H., Bailey, U.-M., Tan, N.Y., Jamaluddin, M.F., Schulz, B.L., 2013. Mixed disulfide formation in vitro between a glycoprotein substrate and yeast oligosaccharyltransferase subunits Ost3p and Ost6p. Biochem. Biophys. Res. Commun. 432, 438–443. https://doi.org/10.1016/j.bbrc.2013.01.128

Mousa, A.A., Cao, S., Aboge, G.O., Terkawi, M.A., El Kirdasy, A., Salama, A., Attia, M., Aboulaila, M., Zhou, M., Kamyingkird, K., Moumouni, P.F.A., Masatani, T., El Aziz, S.A.A., Moussa, W.M., Chahan, B., Fukumoto, S., Nishikawa, Y., El Ballal, S.S., Xuan, X., 2013. Molecular characterization and antigenic properties of a novel Babesia gibsoni glutamic acid-rich protein (BgGARP). Exp. Parasitol. 135, 414–420. https://doi.org/10.1016/j.exppara.2013.08.005

MRAN, 2018. R Packages [WWW Document]. URL https://mran.microsoft.com/package/pheatmap (accessed 6.3.18).

Mukherjee, S., Panda, A., Ghosh, T.C., 2015. Elucidating evolutionary features and functional implications of orphan genes in Leishmania major. Infect. Genet. Evol. 32, 330–337. https://doi.org/10.1016/j.meegid.2015.03.031

Mulhern, T.D., Howlett, G.J., Reid, G.E., Simpson, R.J., McColl, D.J., Anders, R.F., Norton, R.S., 1995. Solution structure of a polypeptide containing four heptad repeat units from a merozoite surface antigen of Plasmodium falciparum. Biochemistry 34, 3479–3491.

Muralidharan, V., Goldberg, D.E., 2013. Asparagine Repeats in Plasmodium falciparum Proteins: Good for Nothing? PLoS Pathog. 9. https://doi.org/10.1371/journal.ppat.1003488

Nandi, T., Dash, D., Ghai, R., B-Rao, C., Kannan, K., Brahmachari, S.K., Ramakrishnan, C., Ramachandran, S., 2003. A novel complexity measure for comparative analysis of protein sequences from complete genomes. J. Biomol. Struct. Dyn. 20, 657–668. https://doi.org/10.1080/07391102.2003.10506882

National Center of Biotechnology Information, 2018. Contamination [WWW Document]. URL https://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/#Sources (accessed 3.15.18).
National Center of Biotechnology Information, 2017. Eukaryotic Annotation Guide [WWW Document]. URL https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission/ (accessed 4.3.17).
Navarre, W.W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S.J., Fang, F.C., 2006. Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella. Science 313, 236–238. https://doi.org/10.1126/science.1128794
Nisbet, R.E.R., McKenzie, J.L., 2016. Transcription of the apicoplast genome. Mol. Biochem. Parasitol. 210, 5–9. https://doi.org/10.1016/j.molbiopara.2016.07.004
Nishizawa, K., Nishizawa, M., Kim, K.S., 1999. Tendency for local repetitiveness in amino acid usages in modern proteins. J. Mol. Biol. 294, 937–953. https://doi.org/10.1006/jmbi.1999.3275
Nørholm, M.H.H., Light, S., Virkki, M.T.I., Elofsson, A., von Heijne, G., Daley, D.O., 2012. Manipulating the genetic code for membrane protein production: what have we learnt so far? Biochim. Biophys. Acta 1818, 1091–1096. https://doi.org/10.1016/j.bbamem.2011.08.018
Notredame, C., Higgins, D.G., Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302 (1), 205–17.
Nudelman, F., Gotliv, B.A., Addadi, L., Weiner, S., 2006. Mollusk shell formation: mapping the distribution of organic matrix components underlying a single aragonitic tablet in nacre. J. Struct. Biol. 153, 176–187. https://doi.org/10.1016/j.jsb.2005.09.009
Oeuvray, C., Bouharoun-Tayoun, H., Gras-Masse, H., Bottius, E., Kaidoh, T., Aikawa, M., Filgueira, M.C., Tartar, A., Druilhe, P., 1994. Merozoite surface protein-3: a malaria protein inducing antibodies that promote Plasmodium falciparum killing by cooperation with blood monocytes. Blood 84, 1594–1602.
Offord, C., 2016. The Pangenome: Are Single Reference Genomes Dead? The Scientist 30.
Ohki, Y., Wenninger-Weinzierl, A., Hruscha, A., Asakawa, K., Kawakami, K., Haass, C., Edbauer, D., Schmid, B., 2017. Glycine-alanine dipeptide repeat protein contributes to toxicity in a zebrafish model of C9orf72 associated neurodegeneration. Mol. Neurodegener. 12, 6. https://doi.org/10.1186/s13024-016-0146-8
Ohno, S., Epplen, J.T., 1983. The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc. Natl. Acad. Sci. 80, 3391–3395.
Okombo, J., Abdi, A.I., Kiara, S.M., Mwai, L., Pole, L., Sutherland, C.J., Nzila, A., Ochola-Oyier, L.I., 2013. Repeat Polymorphisms in the Low-Complexity Regions of Plasmodium falciparum ABC Transporters and Associations with In Vitro Antimalarial Responses. Antimicrob. Agents Chemother. 57, 6196–6204. https://doi.org/10.1128/AAC.01465-13
Ooi, H.S., Kwo, C.Y., Wildpaner, M., Sirota, F.L., Eisenhaber, B., Maurer-Stroh, S., Wong, W.C., Schleiffer, A., Eisenhaber, F., Schneider, G., 2009. ANNIE: integrated de novo protein sequence annotation. Nucleic Acids Res. 37, W435–W440. https://doi.org/10.1093/nar/gkp254
Otto, T. D., Böhme, U., Jackson, A.P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W.A.M., Religa, A.A., Robertson, L., Sanders, M., Ogun, S.A., Cunningham, D., Erhart, A., Billker, O., Khan, S.M., Stunnenberg, H.G., Langhorne, J., Holder, A.A., Waters, A.P., Newbold, C.I., Pain, A., Berriman, M., Janse, C.J., 2014. A comprehensive

evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. 12, 86. https://doi.org/10.1186/s12915-014-0086-0

Otto, Thomas D., Rayner, J.C., Böhme, U., Pain, A., Spottiswoode, N., Sanders, M., Quail, M., Ollomo, B., Renaud, F., Thomas, A.W., Prugnolle, F., Conway, D.J., Newbold, C., Berriman, M., 2014. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat. Commun. 5, 4754. https://doi.org/10.1038/ncomms5754

Outlaw, D.C., Ricklefs, R.E., 2011. Rerooting the evolutionary tree of malaria parasites. Proc. Natl. Acad. Sci. 108, 13183–13187. https://doi.org/10.1073/pnas.1109153108

Ouzounis, C., Kyrpides, N., 1996. The emergence of major cellular processes in evolution. FEBS Lett. 390, 119–123.

Pain, A., Böhme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E.M., Aslett, M.A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T.J., Cherevach, I., Chillingworth, T., Clark, T.G., Galinski, M.R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C.S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I.M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T.J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A.R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M.A., Cowman, A.F., Turner, C.M.R., Rajandream, M.A., Kocken, C.H.M., Thomas, A.W., Newbold, C.I., Barrell, B.G., Berriman, M., 2008. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature 455, 799–803. https://doi.org/10.1038/nature07306

Pattaradilokrat, S., Sawaswong, V., Simpalipan, P., Kaewthamasorn, M., Siripoon, N., Harnyuttanakorn, P., 2016. Genetic diversity of the merozoite surface protein-3 gene in Plasmodium falciparum populations in Thailand. Malar. J. 15, 517. https://doi.org/10.1186/s12936-016-1566-1

Peel, S.A., 2001. The ABC transporter genes of Plasmodium falciparum and drug resistance. Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother. 4, 66–74. https://doi.org/10.1054/drup.2001.0183

Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., Kurgan, L., 2015. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci. CMLS 72, 137–151. https://doi.org/10.1007/s00018-014-1661-9

Pentelute, B.L., Gates, Z.P., Tereshko, V., Dashnau, J.L., Vanderkooi, J.M., Kossiakoff, A.A., Kent, S.B., 2008. X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. J. Am. Chem. Soc. 130, 9695–9701. https://doi.org/10.1021/ja8013538

Perkins, S.L., 2000. Species concepts and malaria parasites: detecting a cryptic species of Plasmodium. Proc. R. Soc. B Biol. Sci. 267, 2345–2350.

Perkins, S.L., Schall, J.J., 2002. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. J. Parasitol. 88, 972–978. https://doi.org/10.1645/0022-3395(2002)088[0972:AMPOMP]2.0.CO;2

Pfeffer, S., Dudek, J., Gogala, M., Schorr, S., Linxweiler, J., Lang, S., Becker, T., Beckmann, R., Zimmermann, R., Förster, F., 2014. Structure of the mammalian oligosaccharyl-transferase complex in the native ER protein translocon. Nat. Commun. 5, 3072. https://doi.org/10.1038/ncomms4072

Pick, C., Ebersberger, I., Spielmann, T., Bruchhaus, I., Burmester, T., 2011. Phylogenomic analyses of malaria parasites and evolution of their exported proteins. BMC Evol. Biol. 11, 167. https://doi.org/10.1186/1471-2148-11-167

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

Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Ouzounis, C.A., 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics 16, 915–922.

Promponas, V.J., Katsani, K.R., Blencowe, B.J., Ouzounis, C.A., 2016. Sequence evidence for common ancestry of eukaryotic endomembrane coatomers. Sci. Rep. 6, 22311. https://doi.org/10.1038/srep22311

Psomopoulos, Fotis E., Siarkou, V.I., Papanikolaou, N., Iliopoulos, I., Tsaftaris, A.S., Promponas, V.J., Ouzounis, C.A., 2012. The Chlamydiales Pangenome Revisited: Structural Stability and Functional Coherence. Genes 3, 291–319. https://doi.org/10.3390/genes3020291

Psomopoulos, F. E., Siarkou, V.I., Papanikolaou, N., Illiopoulos, I., Tsaftaris, A.S., Promponas, V.J., Ouzounis, C.A., 2012. The Chlamydiales Pangenome Revisited: Structural Stability and Functional Coherence. Genes Basel 3, 291–319.

Qari, S.H., Shi, Y.P., Pieniazek, N.J., Collins, W.E., Lal, A.A., 1996. Phylogenetic relationship among the malaria parasites based on small subunit rRNA gene sequences: monophyletic nature of the human malaria parasite, Plasmodium falciparum. Mol. Phylogenet. Evol. 6, 157–165.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ralph, S.A., Van Dooren, G.G., Waller, R.F., Crawford, M.J., Fraunholz, M.J., Foth, B.J., Tonkin, C.J., Roos, D.S., McFadden, G.I., 2004. Tropical infectious diseases: metabolic maps and functions of the Plasmodium falciparum apicoplast. Nat. Rev. Microbiol. 2, 203–216. https://doi.org/10.1038/nrmicro843

Reid, A.J., Vermont, S.J., Cotton, J.A., Harris, D., Hill-Cawthorne, G.A., Könen-Waisman, S., Latham, S.M., Mourier, T., Norton, R., Quail, M.A., Sanders, M., Shanmugam, D., Sohal, A., Wasmuth, J.D., Brunk, B., Grigg, M.E., Howard, J.C., Parkinson, J., Roos, D.S., Trees, A.J., Berriman, M., Pain, A., Wastling, J.M., 2012. Comparative Genomics of the Apicomplexan Parasites Toxoplasma gondii and Neospora caninum: Coccidia Differing in Host Range and Transmission Strategy. PLOS Pathog. 8, e1002567. https://doi.org/10.1371/journal.ppat.1002567

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J., Jones, C.D., 2013. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLOS Genet. 9, e1003860. https://doi.org/10.1371/journal.pgen.1003860

Reiss, G., te Heesen, S., Gilmore, R., Zufferey, R., Aebi, M., 1997. A specific screen for oligosaccharyltransferase mutations identifies the 9 kDa OST5 protein required for optimal activity in vivo and in vitro. EMBO J. 16, 1164–1172. https://doi.org/10.1093/emboj/16.6.1164

Reynolds, S.M., Käll, L., Riffle, M.E., Bilmes, J.A., Noble, W.S., 2008. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput. Biol. 4, e1000213. https://doi.org/10.1371/journal.pcbi.1000213

Ribeiro, D.M., Briere, G., Bely, B., Spinelli, L., Brun, C., 2018. MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1039

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., Dunker, A.K., 2000. Sequence complexity of disordered protein. Proteins Struct. Funct. Bioinforma. 42, 38–48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3

Römisch, K., 2005. Protein Targeting from Malaria Parasites to Host Erythrocytes. Traffic 6, 706–709. https://doi.org/10.1111/j.1600-0854.2005.00310.x

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma. Oxf. Engl. 19, 1572–1574.

RStudio Team, 2015. RStudio: Integrated Development for R. RStudio Team (2015), RStudio, Inc., Boston, MA.

Rutledge, G.G., Böhme, U., Sanders, M., Reid, A.J., Cotton, J.A., Maiga-Ascofare, O., Djimdé, A.A., Apinjoh, T.O., Amenga-Etego, L., Manske, M., Barnwell, J.W., Renaud, F., Ollomo, B., Prugnolle, F., Anstey, N.M., Auburn, S., Price, R.N., McCarthy, J.S., Kwiatkowski, D.P., Newbold, C.I., Berriman, M., Otto, T.D., 2017. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature 542, 101–104. https://doi.org/10.1038/nature21038

Sakamoto, H., Takeo, S., Maier, A.G., Sattabongkot, J., Cowman, A.F., Tsuboi, T., 2012. Antibodies against a Plasmodium falciparum antigen PfMSPDBL1 inhibit merozoite invasion into human erythrocytes. Vaccine 30, 1972–1980. https://doi.org/10.1016/j.vaccine.2012.01.010

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12, 87. https://doi.org/10.1186/s12915-014-0087-z

Samuelson, J., Banerjee, S., Magnelli, P., Cui, J., Kelleher, D.J., Gilmore, R., Robbins, P.W., 2005. The diversity of dolichol-linked precursors to Asn-linked glycans likely results from secondary loss of sets of glycosyltransferases. Proc. Natl. Acad. Sci. U. S. A. 102, 1548–1553. https://doi.org/10.1073/pnas.0409460102

Samuelson, J., Robbins, P.W., 2015. Effects of N-glycan precursor length diversity on quality control of protein folding and on protein glycosylation. Semin. Cell Dev. Biol. 41, 121–128. https://doi.org/10.1016/j.semcdb.2014.11.008

Saqi, M., 1995. An analysis of structural instances of low complexity sequence segments. Protein Eng. 8, 1069–1073.

Sargeant, T.J., Marti, M., Caler, E., Carlton, J.M., Simpson, K., Speed, T.P., Cowman, A.F., 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. Genome Biol. 7, R12. https://doi.org/10.1186/gb-2006-7-2-r12

Sato, S., Sesay, A.K., Holder, A.A., 2013. The unique structure of the apicoplast genome of the rodent malaria parasite Plasmodium chabaudi chabaudi. PloS One 8, e61778. https://doi.org/10.1371/journal.pone.0061778

Sayers, E., Miller, V., 2010. The E-utility Web Service (SOAP). National Center for Biotechnology Information (US), Bethesda (MD).

Schaer, J., Perkins, S.L., Decher, J., Leendertz, F.H., Fahr, J., Weber, N., Matuschewski, K., 2013. High diversity of West African bat malaria parasites and a tight link with rodent Plasmodium taxa. Proc. Natl. Acad. Sci. U. S. A. 110, 17415–17419. https://doi.org/10.1073/pnas.1311016110

Schnell, S., Fortunato, S., Roy, S., 2007. Is the intrinsic disorder of proteins the cause of the scale-free architecture of protein-protein interaction networks? Proteomics 7, 961–964. https://doi.org/10.1002/pmic.200600455

Schwartz, R.L., Phoenix, T., Foy, B.D., 2005. In the World of Regular Expressions, in: Learning Perl. O'Reilly Media.

Schwarz, F., Aebi, M., 2011. Mechanisms and principles of N-linked protein glycosylation. Curr. Opin. Struct. Biol., Carbohydrates and glycoconjugates/Biophysical methods 21, 576–582. https://doi.org/10.1016/j.sbi.2011.08.005

Shibatani, T., David, L.L., McCormack, A.L., Frueh, K., Skach, W.R., 2005. Proteomic analysis of mammalian oligosaccharyltransferase reveals multiple subcomplexes that contain Sec61, TRAP, and two potential new subunits. Biochemistry 44, 5982–5992. https://doi.org/10.1021/bi047328f

Shin, S.W., Kim, S.M., 2005. A new algorithm for detecting low-complexity regions in protein sequences. Bioinformatics 21, 160–170. https://doi.org/10.1093/bioinformatics/bth497

Shrimal, S., Cherepanova, N.A., Gilmore, R., 2017. DC2 and KCP2 mediate the interaction between the oligosaccharyltransferase and the ER translocon. J. Cell Biol. 216, 3625–3638. https://doi.org/10.1083/jcb.201702159

Shrimal, S., Gilmore, R., 2018. Oligosaccharyltransferase structures provide novel insight into the mechanism of asparagine-linked glycosylation in prokaryotic and eukaryotic cells. Glycobiology. https://doi.org/10.1093/glycob/cwy093

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539–539. https://doi.org/10.1038/msb.2011.75

Sigaux, F., 2000. [Cancer genome or the development of molecular portraits of tumors]. Bull. Acad. Natl. Med. 184, 1441–1447; discussion 1448-1449.

Sim, K.L., Creamer, T.P., 2004. Protein simple sequence conservation. Proteins 54, 629–638. https://doi.org/10.1002/prot.10623

Simon, M., Hancock, J.M., 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol. 10, R59. https://doi.org/10.1186/gb-2009-10-6-r59

Singh, B., Kim Sung, L., Matusop, A., Radhakrishnan, A., Shamsul, S.S.G., Cox-Singh, J., Thomas, A., Conway, D.J., 2004. A large focus of naturally acquired Plasmodium knowlesi infections in human beings. Lancet Lond. Engl. 363, 1017–1024. https://doi.org/10.1016/S0140-6736(04)15836-4

Singh, G.P., Chandra, B.R., Bhattacharya, A., Akhouri, R.R., Singh, S.K., Sharma, A., 2004. Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum. Mol. Biochem. Parasitol. 137, 307–319. https://doi.org/10.1016/j.molbiopara.2004.05.016

Singh, S., Soe, S., Weisman, S., Barnwell, J.W., Pérignon, J.L., Druilhe, P., 2009. A Conserved Multi-Gene Family Induces Cross-Reactive Antibodies Effective in Defense against Plasmodium falciparum. PLoS ONE 4. https://doi.org/10.1371/journal.pone.0005410

Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., Krogh, A., 2001. On the total number of genes and their length distribution in complete microbial genomes. Trends Genet. 17, 425–428. https://doi.org/10.1016/S0168-9525(01)02372-1

Skrabana, R., Skrabanova, M., Csokova, N., Sevcik, J., Novak, M., 2006. Intrinsically disordered tau protein in Alzheimer's tangles: a coincidence or a rule? Bratisl. Lek. Listy 107, 354–358.

Söding, J., Remmert, M., 2011. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr. Opin. Struct. Biol. 21, 404–411. https://doi.org/10.1016/j.sbi.2011.03.005

Solovyev, V., Kosarev, P., Seledsov, I., Vorobyev, D., 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7 Suppl 1, S10.1-12. https://doi.org/10.1186/gb-2006-7-s1-s10

Solovyev, V.V., 2007. Statistical approaches in Eukaryotic gene prediction., in: Balding, D., Cannings, C., Bishop, M. (Eds.), In Handbook of Statistical Genetics. Wiley-Interscience, p. 1616p.

Spielmann, T., Beck, H.P., 2000. Analysis of stage-specific transcription in Plasmodium falciparum reveals a set of genes exclusively transcribed in ring stage parasites. Mol. Biochem. Parasitol. 111, 453–458. https://doi.org/10.1016/S0166-6851(00)00333-9

Spirig, U., Bodmer, D., Wacker, M., Burda, P., Aebi, M., 2005. The 3.4-kDa Ost4 protein is required for the assembly of two distinct oligosaccharyltransferase complexes in yeast. Glycobiology 15, 1396–1406. https://doi.org/10.1093/glycob/cwj025

Spurrier, J.D., 2003. On the null distribution of the Kruskal–Wallis statistic. J. Nonparametric Stat. 15, 685–691. https://doi.org/10.1080/10485250310001634719

Stamnes, M.A., Rutherford, S.L., Zuker, C.S., 1992. Cyclophilins: a new family of proteins involved in intracellular folding. Trends Cell Biol. 2, 272–276. https://doi.org/10.1016/0962-8924(92)90200-7

Steinley, D., 2004. Properties of the Hubert-Arable Adjusted Rand Index. Psychol. Methods 9, 386–396.

Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A., Wellems, T.E., 1995. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82, 89–100.

Subramanyam, M.B., Gnanamani, M., Ramachandran, S., 2006. Simple sequence proteins in prokaryotic proteomes. BMC Genomics 7, 141. https://doi.org/10.1186/1471-2164-7-141

Ta, T.H., Hisam, S., Lanza, M., Jiram, A.I., Ismail, N., Rubio, J.M., 2014. First case of a naturally acquired human infection with Plasmodium cynomolgi. Malar. J. 13.

Tachibana, S.-I., Sullivan, S.A., Kawai, S., Nakamura, S., Kim, H.R., Goto, N., Arisue, N., Palacpac, N.M.Q., Honma, H., Yagi, M., Tougan, T., Katakai, Y., Kaneko, O., Mita, T., Kita, K., Yasutomi, Y., Sutton, P.L., Shakhbatyan, R., Horii, T., Yasunaga, T., Barnwell, J.W., Escalante, A.A., Carlton, J.M., Tanabe, K., 2012. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. Nat. Genet. 44, 1051–1055. https://doi.org/10.1038/ng.2375

Tamana, S., Kirmitzoglou, I., Promponas, V.J., 2012. Sequence features of Compositionally Biased regions in three dimensional protein structures, in: Bioinformatics & Bioengineering (BIBE). Presented at the IEEE 12th International Conference on. https://doi.org/10.1109/BIBE.2012.6399687

Tamana, S., Promponas, V.J., 2018. An updated view of the oligosaccharyltransferase complex in Plasmodium. Submitt. Glycobiol.

Tautz, D., Domazet-Lošo, T., 2011. The evolutionary origin of orphan genes. Nat. Rev. Genet. 12, 692–702. https://doi.org/10.1038/nrg3053

Tautz, D., Trick, M., Dover, G.A., 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature 322, 652–656. https://doi.org/10.1038/322652a0

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. U. S. A. 102, 13950–13955. https://doi.org/10.1073/pnas.0506758102

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169. https://doi.org/10.1093/nar/gkw1099

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A.K., Daughdrill, G.W., Uversky, V.N., 2013. The alphabet of intrinsic disorder. Intrinsically Disord. Proteins 1. https://doi.org/10.4161/idp.24360

Toll-Riera, M., Radó-Trilla, N., Martys, F., Albà, M.M., 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol. Biol. Evol. 29, 883–886. https://doi.org/10.1093/molbev/msr263

Tompa, P., Kovacs, D., 2010. Intrinsically disordered chaperones in plants and animals. Biochem. Cell Biol. Biochim. Biol. Cell. 88, 167–174. https://doi.org/10.1139/o09-163

Tonkin, C.J., Pearce, J.A., McFadden, G.I., Cowman, A.F., 2006. Protein targeting to destinations of the secretory pathway in the malaria parasite Plasmodium falciparum. Curr. Opin. Microbiol., Host microbe interactions: fungi/Host microbe interactions: parasites/Host microbe interactions: viruses 9, 381–387. https://doi.org/10.1016/j.mib.2006.06.015

Troshin, P.V., Procter, J.B., Barton, G.J., 2011. Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA. Bioinformatics 27, 2001–2002.

Tsirigos, K.D., Peters, C., Shu, N., Käll, L., Elofsson, A., 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 43, W401-407. https://doi.org/10.1093/nar/gkv485

Tusnády, G.E., Dosztányi, Z., Simon, I., 2005. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinforma. Oxf. Engl. 21, 1276–1277. https://doi.org/10.1093/bioinformatics/bti121

Tyagi, S., Sharma, M., Das, A., 2011. Comparative genomic analysis of simple sequence repeats in three Plasmodium species. Parasitol. Res. 108, 451–458. https://doi.org/10.1007/s00436-010-2086-5

Uversky, V.N., 2013. The alphabet of intrinsic disorder. Intrinsically Disord. Proteins 1. https://doi.org/10.4161/idp.24684

Uversky, V.N., Dunker, A.K., 2010. Understanding protein non-folding. Biochim. Biophys. Acta 1804, 1231–1264. https://doi.org/10.1016/j.bbapap.2010.01.017

Uversky, V.N., Gillespie, J.R., Fink, A.L., 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41, 415–427.

Van Dongen, S., 2000. A cluster algorithm for graphs. Tech. Rep. -R0010 Natl. Res. Inst. Math. Comput. Sci. Neth. Amst.

Vanin, E.F., 1985. Processed Pseudogenes: Characteristics and Evolution. Annu. Rev. Genet. 19, 253–272. https://doi.org/10.1146/annurev.ge.19.120185.001345

Vaughan, A., Chiu, S.-Y., Ramasamy, G., Li, L., Gardner, M.J., Tarun, A.S., Kappe, S.H.I., Peng, X., 2008. Assessment and improvement of the Plasmodium yoelii yoelii genome annotation through comparative analysis. Bioinforma. Oxf. Engl. 24, i383-389. https://doi.org/10.1093/bioinformatics/btn140

Veiga, M.I., Ferreira, P.E., Jörnhagen, L., Malmberg, M., Kone, A., Schmidt, B.A., Petzold, M., Björkman, A., Nosten, F., Gil, J.P., 2011. Novel Polymorphisms in Plasmodium falciparum ABC Transporter Genes Are Associated with Major ACT Antimalarial Drug Resistance. PLOS ONE 6, e20212. https://doi.org/10.1371/journal.pone.0020212

Vernikos, G., Medini, D., Riley, D.R., Tettelin, H., 2015. Ten years of pan-genome analyses. Curr. Opin. Microbiol., Host–microbe interactions: bacteria • Genomics 23, 148–154. https://doi.org/10.1016/j.mib.2014.11.016

Verra, F., Hughes, A.L., 1999. Biased amino acid composition in repeat regions of Plasmodium antigens. Mol. Biol. Evol. 16, 627–633. https://doi.org/10.1093/oxfordjournals.molbev.a026145

Verster, A.J., Styles, E.B., Mateo, A., Derry, W.B., Andrews, B.J., Fraser, A.G., 2017. Taxonomically Restricted Genes with Essential Functions Frequently Play Roles in Chromosome Segregation in Caenorhabditis elegans and Saccharomyces cerevisiae. G3 Genes Genomes Genet. 7, 3337–3347. https://doi.org/10.1534/g3.117.300193

Vijay-Kumar, M., Aitken, J.D., Carvalho, F.A., Cullender, T.C., Mwangi, S., Srinivasan, S., Sitaraman, S.V., Knight, R., Ley, R.E., Gewirtz, A.T., 2010. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. Science 328, 228–231. https://doi.org/10.1126/science.1179721

Vishnoi, A., Kryazhimskiy, S., Bazykin, G.A., Hannenhalli, S., Plotkin, J.B., 2010. Young proteins experience more variable selection pressures than old proteins. Genome Res. 20, 1574–1581. https://doi.org/10.1101/gr.109595.110

von Itzstein, M., Plebanski, M., Cooke, B.M., Coppel, R.L., 2008. Hot, sweet and sticky: the glycobiology of Plasmodium falciparum. Trends Parasitol. 24, 210–218. https://doi.org/10.1016/j.pt.2008.02.007

Walker, F.O., 2007. Huntington's disease. Lancet Lond. Engl. 369, 218–228. https://doi.org/10.1016/S0140-6736(07)60111-1

Waller, R.F., Reed, M.B., Cowman, A.F., McFadden, G.I., 2000. Protein trafficking to the plastid of Plasmodium falciparum is via the secretory pathway. EMBO J. 19, 1794–1802. https://doi.org/10.1093/emboj/19.8.1794

Wampler, J.E., 2010. The 20 Amino Acids: hydrophobic, hydrophilic, polar and charged amino acids [WWW Document]. Struct. Bioinforma. Protein Crystallogr. Seq. Anal. Homol. Model. URL https://proteinstructures.com/Structure/Structure/amino-acids.html (accessed 5.26.18).

Wan, H., Li, L., Federhen, S., Wootton, J.C., 2003. Discovering simple regions in biological sequences associated with scoring schemes. J. Comput. Biol. J. Comput. Mol. Cell Biol. 10, 171–185. https://doi.org/10.1089/106652703321825955

Wang, G., Dunbrack, R.L.J., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591. https://doi.org/10.1093/bioinformatics/btg224

Wang, J., Zheng, W., Liu, F., Wang, Y., He, Y., Zheng, L., Fan, Q., Luo, E., Cao, Y., Cui, L., 2017. Characterization of Pb51 in Plasmodium berghei as a malaria vaccine candidate targeting both asexual erythrocytic proliferation and transmission. Malar. J. 16, 458. https://doi.org/10.1186/s12936-017-2107-2

Warncke, J.D., Vakonakis, I., Beck, H.-P., 2016. Plasmodium Helical Interspersed Subtelomeric (PHIST) Proteins, at the Center of Host Cell Remodeling. Microbiol. Mol. Biol. Rev. MMBR 80, 905–927. https://doi.org/10.1128/MMBR.00014-16

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191.

Waters, A.P., Higgins, D.G., McCutchan, T.F., 1993. Evolutionary relatedness of some primate models of Plasmodium. Mol. Biol. Evol. 10, 914–923. https://doi.org/10.1093/oxfordjournals.molbev.a040038

Waters, A.P., Higgins, D.G., McCutchan, T.F., 1991. Plasmodium falciparum appears to have arisen as a result of lateral transfer between avian and human hosts. Proc. Natl. Acad. Sci. U. S. A. 88, 3140–3144.

Wei, L., Liu, Y., Dubchak, I., Shon, J., Park, J., 2002. Comparative genomics approaches to study organism similarities and differences. J. Biomed. Inform. 35, 142–150.

Weiner, S., 1979. Aspartic acid-rich proteins: major components of the soluble organic matrix of mollusk shells. Calcif. Tissue Int. 29, 163–167.

Weiss, C., Bertalan, I., Johanningmeier, U., 2012. Effects of rare codon clusters on the expression of a high-turnover chloroplast protein in Chlamydomonas reinhardtii. J. Biotechnol. 160, 105–111. https://doi.org/10.1016/j.jbiotec.2012.04.008

Weiss, S., Amir, A., Hyde, E.R., Metcalf, J.L., Song, S.J., Knight, R., 2014. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 15. https://doi.org/10.1186/s13059-014-0564-2

Wikipedia, the free encyclopedia, 2018. Box plot. Wikipedia.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. Biom. Bull. 1, 80–83.

Wild, R., Kowal, J., Eyring, J., Ngwa, E.M., Aebi, M., Locher, K.P., 2018. Structure of the yeast oligosaccharyltransferase complex gives insight into eukaryotic N-glycosylation. Science 359, 545–550. https://doi.org/10.1126/science.aar5140

Williams, R.M., Obradovi, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J., Dunker, A.K., 2001. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 89–100.

Wilson, B.A., Foy, S.G., Neme, R., Masel, J., 2017. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis ofDe NovoGene Birth. Nat. Ecol. Evol. 1, 0146–0146. https://doi.org/10.1038/s41559-017-0146

Wilson, G.A., Bertrand, N., Patel, Y., Hughes, J.B., Feil, E.J., Field, D., 2005. Orphans as taxonomically restricted and ecologically important genes. Microbiology 151, 2499–2501. https://doi.org/10.1099/mic.0.28146-0

Wilson, G.A., Feil, E.J., Lilley, A.K., Field, D., 2007. Large-Scale Comparative Genomic Ranking of Taxonomically Restricted Genes (TRGs) in Bacterial and Archaeal Genomes. PLOS ONE 2, e324. https://doi.org/10.1371/journal.pone.0000324

Winter, G., Kawai, S., Haeggström, M., Kaneko, O., Von Euler, A., Kawazu, S.-I., Palm, D., Fernandez, V., Wahlgren, M., 2005. SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. J. Exp. Medicat. 201, 1853–63.

Wise, M.J., 2001. 0j.py: a software tool for low complexity proteins and protein domains. Bioinforma. Oxf. Engl. 17 Suppl 1, S288-295.
Wootton, J.C., 1994. Sequences with 'unusual' amino acid compositions. Curr. Opin. Struct. Biol. 4, 413–421. https://doi.org/10.1016/S0959-440X(94)90111-2
Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17, 149–163. https://doi.org/10.1016/0097-8485(93)85006-X
World Health Organization, 2018. Overview of malaria treatment [WWW Document]. WHO. URL http://www.who.int/malaria/areas/treatment/overview/en/ (accessed 5.20.18).
World Health Organization, 2017. Key points: World malaria report 2017 [WWW Document]. WHO. URL http://www.who.int/malaria/media/world-malaria-report-2017/en/ (accessed 3.18.18).
Wyrick, P.B., 2000. Intracellular survival by Chlamydia. Cell. Microbiol. 2, 275–282. https://doi.org/10.1046/j.1462-5822.2000.00059.x
Ye, J.Z.-S., Donigian, J.R., Van Overbeek, M., Loayza, D., Luo, Y., Krutchinsky, A.N., Chait, B.T., De Lange, T., 2004. TIN2 binds TRF1 and TRF2 simultaneously and stabilizes the TRF2 complex on telomeres. J. Biol. Chem. 279, 47264–47271. https://doi.org/10.1074/jbc.M409047200
Zhang, Y., Stec, B., Godzik, A., 2007. Between order and disorder in protein structures – analysis of "dual personality" fragments in proteins. Struct. Lond. Engl. 1993 15, 1141–1147. https://doi.org/10.1016/j.str.2007.07.012
Zhang, Z.C., Chook, Y.M., 2012. Structural and energetic basis of ALS-causing mutations in the atypical proline-tyrosine nuclear localization signal of the Fused in Sarcoma protein (FUS). Proc. Natl. Acad. Sci. U. S. A. 109, 12017–12021. https://doi.org/10.1073/pnas.1207247109
Zheng, Y., Anton, B.P., Roberts, R.J., Kasif, S., 2005. Phylogenetic detection of conserved gene clusters in microbial genomes. BMC Bioinformatics 6, 243. https://doi.org/10.1186/1471-2105-6-243
Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., Wang, W., 2008. On the origin of new genes in Drosophila. Genome Res. 18, 1446–1455. https://doi.org/10.1101/gr.076588.108
Zhu, L., Mok, S., Imwong, M., Jaidee, A., Russell, B., Nosten, F., Day, N.P., White, N.J., Preiser, P.R., Bozdech, Z., 2016. New insights into the Plasmodium vivax transcriptome using RNA-Seq. Sci. Rep. 6, 20498. https://doi.org/10.1038/srep20498
Zilversmit, M., Perkins, S., 2008. Plasmodium. Malaria Parasites. Tree Life Web Proj. Version5.
Zilversmit, M.M., Pattaradilokrat, S., Su, X., 2016. Comparative and functional genomics of malaria parasites, in: Advances in Malaria Research. Wiley-Blackwell, pp. 125–148. https://doi.org/10.1002/9781118493816.ch5
Zilversmit, M.M., Volkman, S.K., DePristo, M.A., Wirth, D.F., Awadalla, P., Hartl, D.L., 2010. Low-Complexity Regions in Plasmodium falciparum: Missing Links in the Evolution of an Extreme Genome. Mol. Biol. Evol. 27, 2198–2209. https://doi.org/10.1093/molbev/msq108
Zubkov, S., Lennarz, W.J., Mohanty, S., 2004. Structural basis for the function of a minimembrane protein subunit of yeast oligosaccharyltransferase. Proc. Natl. Acad. Sci. U. S. A. 101, 3821–3826. https://doi.org/10.1073/pnas.0400512101

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G.I., Schneider, G., 2001. Deciphering apicoplast targeting signals – feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. Gene 280, 19–26. https://doi.org/10.1016/S0378-1119(01)00776-4

# Appendix I

## Comparative genomics pipeline supplementary materials

- **Datasets**
  - PlasmoDB genomes
  - Pangenome
  - CAST
  - BLAST database files
    - Masked
    - Unmasked
  - MCL input files
- **Results**
  - BLAST alignments
  - MCL output files
  - Pfam domains
  - Functional annotations
- **Source codes**
  - Pangenome analysis

## Supplementary Tables

**Table S- 1:** The results of Wilcoxon Rank Sum test on cluster sequence lengths. Diagonal elements; No-available computations between the same MCL run. Red: Non-significant p-values. The significance level is set to 99.5%.

| | MCL1 | MCL2 | MCL3 | MCL4 | MCL5 | MCL6 | MCL7 | MCL8 | MCL9 | MCL10 | MCL11 | MCL12 | MCL13 | MCL14 | MCL15 | MCL16 | MCL17 | MCL18 | MCL19 | MCL20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MCL1** | N/A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.11 | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 |
| **MCL2** | | N/A | 0.22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.57 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.47 | 0 |
| **MCL3** | | | N/A | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0.07 | 0 | 0 | 0.07 | 0 | 0 | 0 | 0.17 | 0.60 | 0 |
| **MCL4** | | | | N/A | 0.11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **MCL5** | | | | | N/A | 0 | 0 | 0 | 0 | 0.16 | 0.45 | 0 | 0 | 0.42 | 0.91 | 0 | 0 | 0.23 | 0 | 0 |
| **MCL6** | | | | | | N/A | 0.56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **MCL7** | | | | | | | N/A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **MCL8** | | | | | | | | N/A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **MCL9** | | | | | | | | | N/A | 0 | 0 | 0 | 0.67 | 0 | 0 | 0 | 0.54 | 0 | 0 | 0 |
| **MCL10** | | | | | | | | | | N/A | 0.50 | 0 | 0 | 0 | 0.19 | 0 | 0 | 0 | 0.19 | 0 |
| **MCL11** | | | | | | | | | | | N/A | 0 | 0 | 0.95 | 0.51 | 0 | 0 | 0.66 | 0.19 | 0 |
| **MCL12** | | | | | | | | | | | | N/A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **MCL13** | | | | | | | | | | | | | N/A | 0 | 0 | 0 | 0.30 | 0 | 0 | 0 |
| **MCL14** | | | | | | | | | | | | | | N/A | 0.49 | 0 | 0 | 0 | 0.20 | 0 |
| **MCL15** | | | | | | | | | | | | | | | N/A | 0 | 0 | 0.27 | 0.05 | 0 |
| **MCL16** | | | | | | | | | | | | | | | | N/A | 0 | 0 | 0 | 0 |
| **MCL17** | | | | | | | | | | | | | | | | | N/A | 0 | 0 | 0 |
| **MCL18** | | | | | | | | | | | | | | | | | | N/A | 0 | 0 |
| **MCL19** | | | | | | | | | | | | | | | | | | | N/A | 0 |
| **MCL20** | | | | | | | | | | | | | | | | | | | | N/A |

**Table S-2:** The results of Wilcoxon Rank Sum test using the average scores of the Pfam domains. Red: Significant scores. The significance level is set to 99.5%.

| | MCL1 | MCL2 | MCL3 | MCL4 | MCL5 | MCL6 | MCL7 | MCL8 | MCL9 | MCL10 | MCL11 | MCL12 | MCL13 | MCL14 | MCL15 | MCL16 | MCL17 | MCL18 | MCL19 | MCL20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MCL1** | N/A | 0.17 | 0.15 | 0.00 | 0.15 | 0.01 | 0.01 | 0.00 | 0.70 | 0.17 | 0.23 | 0.00 | 0.66 | 0.21 | 0.27 | 0.00 | 0.87 | 0.17 | 0.13 | 0.00 |
| **MCL2** | 0.17 | N/A | 0.95 | 0.10 | 0.95 | 0.17 | 0.21 | 0.06 | 0.08 | 0.99 | 0.84 | 0.10 | 0.07 | 0.90 | 0.77 | 0.12 | 0.13 | 0.99 | 0.89 | 0.00 |
| **MCL3** | 0.15 | 0.95 | N/A | 0.12 | 1.00 | 0.19 | 0.23 | 0.06 | 0.07 | 0.95 | 0.80 | 0.11 | 0.06 | 0.85 | 0.72 | 0.14 | 0.11 | 0.94 | 0.94 | 0.00 |
| **MCL4** | 0.00 | 0.10 | 0.12 | N/A | 0.11 | 0.74 | 0.66 | 0.83 | 0.00 | 0.10 | 0.07 | 0.99 | 0.00 | 0.08 | 0.05 | 0.90 | 0.00 | 0.10 | 0.13 | 0.09 |
| **MCL5** | 0.15 | 0.95 | 1.00 | 0.11 | N/A | 0.18 | 0.22 | 0.06 | 0.06 | 0.95 | 0.79 | 0.11 | 0.05 | 0.85 | 0.72 | 0.13 | 0.10 | 0.94 | 0.94 | 0.00 |
| **MCL6** | 0.01 | 0.17 | 0.19 | 0.74 | 0.18 | N/A | 0.91 | 0.57 | 0.00 | 0.16 | 0.11 | 0.75 | 0.00 | 0.13 | 0.09 | 0.83 | 0.00 | 0.16 | 0.21 | 0.03 |
| **MCL7** | 0.01 | 0.21 | 0.23 | 0.66 | 0.22 | 0.91 | N/A | 0.50 | 0.00 | 0.20 | 0.14 | 0.67 | 0.00 | 0.16 | 0.11 | 0.74 | 0.00 | 0.20 | 0.25 | 0.02 |
| **MCL8** | 0.00 | 0.06 | 0.06 | 0.83 | 0.06 | 0.57 | 0.50 | N/A | 0.00 | 0.05 | 0.03 | 0.81 | 0.00 | 0.04 | 0.03 | 0.73 | 0.00 | 0.05 | 0.07 | 0.11 |
| **MCL9** | 0.70 | 0.08 | 0.07 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | N/A | 0.08 | 0.11 | 0.00 | 0.95 | 0.10 | 0.13 | 0.00 | 0.82 | 0.08 | 0.05 | 0.00 |
| **MCL10** | 0.17 | 0.99 | 0.95 | 0.10 | 0.95 | 0.16 | 0.20 | 0.05 | 0.08 | N/A | 0.85 | 0.10 | 0.07 | 0.90 | 0.77 | 0.12 | 0.12 | 1.00 | 0.89 | 0.00 |
| **MCL11** | 0.23 | 0.84 | 0.80 | 0.07 | 0.79 | 0.11 | 0.14 | 0.03 | 0.11 | 0.85 | N/A | 0.06 | 0.10 | 0.94 | 0.92 | 0.08 | 0.17 | 0.85 | 0.74 | 0.00 |
| **MCL12** | 0.10 | 0.11 | 0.99 | 0.11 | 0.75 | 0.67 | 0.81 | 0.00 | 0.10 | 0.06 | 0.00 | N/A | 0.07 | 0.05 | 0.92 | 0.00 | 0.09 | 0.13 | 0.08 | 0.11 |
| **MCL13** | 0.66 | 0.07 | 0.06 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.95 | 0.07 | 0.10 | 0.00 | N/A | 0.08 | 0.12 | 0.00 | 0.77 | 0.06 | 0.05 | 0.00 |
| **MCL14** | 0.21 | 0.90 | 0.85 | 0.08 | 0.85 | 0.13 | 0.16 | 0.04 | 0.10 | 0.90 | 0.94 | 0.07 | 0.08 | N/A | 0.87 | 0.09 | 0.15 | 0.90 | 0.79 | 0.00 |
| **MCL15** | 0.27 | 0.77 | 0.72 | 0.05 | 0.72 | 0.09 | 0.11 | 0.03 | 0.13 | 0.77 | 0.92 | 0.05 | 0.12 | 0.87 | N/A | 0.06 | 0.20 | 0.77 | 0.66 | 0.00 |
| **MCL16** | 0.00 | 0.12 | 0.14 | 0.90 | 0.13 | 0.83 | 0.74 | 0.73 | 0.00 | 0.12 | 0.08 | 0.92 | 0.00 | 0.09 | 0.06 | N/A | 0.00 | 0.11 | 0.15 | 0.06 |
| **MCL17** | 0.87 | 0.13 | 0.11 | 0.00 | 0.10 | 0.00 | 0.00 | 0.00 | 0.82 | 0.12 | 0.17 | 0.00 | 0.77 | 0.15 | 0.20 | 0.00 | N/A | 0.12 | 0.09 | 0.00 |

| | MCL1 | MCL2 | MCL3 | MCL4 | MCL5 | MCL6 | MCL7 | MCL8 | MCL9 | MCL10 | MCL11 | MCL12 | MCL13 | MCL14 | MCL15 | MCL16 | MCL17 | MCL18 | MCL19 | MCL20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MCL18** | 0.17 | 0.99 | 0.94 | 0.10 | 0.94 | 0.16 | 0.20 | 0.05 | 0.08 | 1.00 | 0.85 | 0.09 | 0.06 | 0.90 | 0.77 | 0.11 | 0.12 | N/A | 0.88 | 0.00 |
| **MCL19** | 0.13 | 0.89 | 0.94 | 0.13 | 0.94 | 0.21 | 0.25 | 0.07 | 0.05 | 0.89 | 0.74 | 0.13 | 0.05 | 0.79 | 0.66 | 0.15 | 0.09 | 0.88 | N/A | 0.00 |
| **MCL20** | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 0.03 | 0.02 | 0.11 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | N/A |

# Appendix II

## Unique genes in malaria parasites: a pangenomic approach

- **Datasets**
  - PlasmoDB genomes
  - NCBI genomes
    - *Plasmodium gallinicaeum* 8A
    - *Plasmodium relictum* SGS1-like
  - *Plasmodium* pangenome
  - *Toxoplasma gondii* ME49
  - CAST
    - *Plasmodium* pangenome
    - *T. gondii ME49* genome
  - BLAST
    - Database
    - Bl2seq
    - R-fraction
    - BBH
- **Results**
  - *Plasmodium* core genome phylogenetic tree
  - Excel files
    - Full summary table of the analysis
    - Taxonomically Restricted Genes presence/absence profiles
  - Unique sequences
  - Statistics
  - Taxonomically Restricted Genes analysis files
  - BLAST output files
    - BLASTP
    - TBLASTN
  - Annotation
  - De novo genes
  - Multiple hits genes
  - Orphans

- Possible contamination
- Species specific genes
- TBLASTN interesting cases

- **Source codes**
  - Phylogenetic tree
  - General
  - Analysis
  - QIPP scores
  - MSA

# Appendix III

## An updated view of the oligosaccharyltransferase complex in Plasmodium

- Supplementary tables
- MSA
- Sequences
- Structure predictions
- TMHMM

### Supplementary text

**Ost4 sequences** – *P. falciparum* fragile strain nilgiri (obtained from PlasmoDB)

Transcript id: AK88_01616-t30_1

Transcript length: 108

Protein length: 35

Genomic length: 306

**Predicted Protein Sequence**
35 aa

MDYELFLVSNILGISIVILIFIFHYLYADEGNDVS

**Predicted RNA/mRNA Sequence (Introns spliced out)**

108 bp

ATGGACTATGAGCTGTTCTTAGTCTCAAACATCCTGGGCATTTCCATAGTCATTTTGATTTTCATTTTTCATTATTTATA

TGCGGATGAGGGGAATGATGTGTCATGA

**Genomic Sequence (Introns highlighted)**

306 bp intron

ATGGACTATGAGCTGTTCTTAGTCTCAAACATCCTGGgtaattttctgcatatatgttgggagatcggcgacaaatgctc

tgcccgctgtttagcttcatagctgcaaagggaatttgcacgaaatccttctcagtgatatcgcagcgggcgcagcgact

gagcgaaggaggtccatccgagtgtaccgcctacacgagtataccgtccatcctacccatgccacccccctgcagGCATT

TCCATAGTCATTTTGATTTTCATTTTTCATTATTTATATGCGGATGAGGGGAATGATGTGTCATGA

**Ost5 sequence** – C. parvum strain Iowa II (obtained from CryptoDB)

**Transcript id:** cgd7_2153-RA

**Transcript length:** 499

**Protein length:** 80

**Genomic length:** 560

**Predicted Protein Sequence**

80 aa

MHSNFKAEPFPLFVSTQSFLPISMILSLISLVFISFFIVYEFRFPRLKRSLKDELIVSFAGSLTLGLGLTFALLSFGLYF

**Predicted RNA/mRNA Sequence (Introns spliced out; UTRs highlighted)**

499 bp utr

ataatctttattattaataattataatcattcaaaccatacattttattaaaatcgaattgcggcaacgctttttatgta
gtaatttttcttgcacaaaagttttaattatattttagggattataatttattgtggaaattaaacaATGCACAGCAACTT
TAAGGCAGAGCCTTTTCCATTATTCGTCTCTACTCAAAGTTTTCTTCCAATATCGATGATTCTCTCGCTTATTTCACTGG
TATTTATATCGTTTTTCATTGTATATGAATTCAGATTCCCAAGATTAAAAAGAAGTTTAAAGGATGAACTAATTGTTTCA
TTTGCAGGATCACTTACTTTAGGCTTAGGGCTCACTTTTGCATTATTATCATTTGGTCTATATTTTTAGgttttagattt
agtttctttaagtactttatctcaagcattactgtaattgtttgaattaagacaaaatttaaaaaaattaactaaagcag
aagatcatagttagagatt

**Genomic Sequence (Introns and UTRs highlighted)**

560 bp intron utr

ataatctttattattaataattataatcattcaaaccatacattttattaaaatcgaattgcggcaacgctttttatgta
gtaatttttcttgcacaaaagttttaattatattttagggattataatttattgtggaaattaaacaATGCACAGCAACTT
TAAGGCAGAGCCTTTTCCATTATTCGTCTCTACTCAAAGTTTTCTTCCAATATCGATGATTCTCTCGCTTATTTgtaagt
ttgattgtatttttttaaaaatttggttaagtattatttatgttttattctccagCACTGGTATTTATATCGTTTTTCAT
TGTATATGAATTCAGATTCCCAAGATTAAAAAGAAGTTTAAAGGATGAACTAATTGTTTCATTTGCAGGATCACTTACTT
TAGGCTTAGGGCTCACTTTTGCATTATTATCATTTGGTCTATATTTTTAGgttttagatttagtttctttaagtacttta
tctcaagcattactgtaattgtttgaattaagacaaaatttaaaaaaattaactaaagcagaagatcatagttagagatt

**Swp1 sequences** – C. parvum strain Iowa II (obtained from CryptoDB)

**Transcript id:** cgd7_5080-RA

**Transcript length:** 2013

**Protein length:** 670

**Genomic length:** 2013

**Predicted Protein Sequence**

670 aa

MGKILVSLILFVFALINKVQTQFVNILPEVPSDITQEIGTCDALLLKLLATNYVPKSETNSLCNIFRNKIQRIDTPLNDL
NCSIFGSILVQCDLKLAITSSEILRQYFMKSTKPTLKELYYLTTILSNLPQEISSVLGRDDIIEDSNEIINDKLSIIKKQ
ISSNQILDYEEISFLFGSISQLIRLGSRNVQHLKMMNTILELQNKNLKFYPIVDLDTFVLNFSLLSKAYFFLRRPVDFED
IPLSLQQMAYSISRIKLRDIKSLTVLNEYIVTRRLFIDSGIYLIQGFQNTNPKNIDENLATVIFCRLDGSPIENLSLDLS
NTNTFKLVGIPNTCEYQLFYNIANQTEPFNLEKVKNDLSKSIPKLKISSRSVLAREANIPILNKKSSNFEITKATIYFGE
KLTLGKSQVKDINTHYMKTLDSFQFSLELQTSLEILNFETVYQYFTAFEIKALTPYKTIGTRTIKLIPAIIEKNNLKLYL
DLTNNDFICSSLEFSIRIIIGNPLKNESSNSSNSNQEILSIMINDERNDTSALSKLCPYKIHPKIADFYPKEEISYKFKE
PRKLPGPLLPYGFSILILLSLFKIIPIWNKLSKIDQGFFTFINNVPFIKVATFISLAISLFIILCYWHTLNIFQFMYIFT
PAICIFFVLLKLSLRNFNYPSFGNEKSKSD

**Predicted RNA/mRNA Sequence (Introns spliced out)**

2013 bp

ATGGGAAAAATACTTGTTTCTTTAATATTATTTGTTTTTGCCTTAATAAATAAGGTACAAACCCAATTCGTTAATATTTT
ACCAGAGGTACCGAGTGATATCACGCAAGAAATAGGTACATGTGATGCATTATTATTAAAATTACTAGCAACAAATTATG
TCCCAAAATCAGAAACTAACTCACTTTGTAACATTTTCAGAAATAAGATTCAAAGGATAGATACTCCATTAAATGATTTA
AATTGCTCAATATTTGGTTCGATACTAGTTCAATGTGATCTCAAATTAGCAATTACTAGTTCTGAAATCTTAAGACAATA
TTTTATGAAATCAACTAAGCCAACCTTAAAAGAGCTATATTACTTAACAACTATACTATCCAATTTACCACAGGAAATTA
GTTCAGTTTTAGGAAGAGACGATATAATAGAAGACTCAAATGAGATTATTAATGATAAGTTAAGTATTATCAAGAAACAA
ATTTCAAGTAATCAAATATTAGATTATGAAGAAATCTCATTTTTATTTGGATCAATTTCTCAGTTAATTCGATTAGGATC
TAGAAATGTACAACATCTTAAAATGATGAATACAATCTTGGAATTGCAAAATAAGAATCTCAAGTTTTATCCTATTGTAG
ACTTAGATACCTTTGTCCTCAACTTCTCACTTTTATCTAAAGCGTACTTCTTCTTAAGAAGACCCGTTGATTTTGAAGAT
ATCCCACTTTCCTTACAGCAAATGGCCTATAGTATTTCAAGAATTAAACTTAGAGATATAAAAAGTTTAACAGTTTTGAA
TGAATATATTGTCACTAGAAGACTGTTTATTGATTCTGGTATTTATTTGATACAAGGTTTCCAAAATACTAATCCCAAAA
ATATTGATGAGAATTTAGCTACAGTTATATTTTGTAGATTAGATGGATCACCAATTGAAAATTTATCTTTGGATTTATCA
AATACTAATACTTTTAAACTTGTAGGTATACCAAATACTTGCGAGTATCAACTTTTTTATAATATTGCAAATCAAACCGA
ACCATTTAATCTTGAGAAAGTGAAGAATGATTTGTCTAAATCTATTCCAAAGCTAAAAATCTCTAGTAGATCAGTCCTTG
CTAGAGAAGCCAATATTCCAATTCTTAATAAGAAAAGCTCCAATTTTGAGATCACAAAAGCCACTATTTACTTTGGTGAG
AAGTTAACATTGGGAAAATCTCAAGTTAAGGATATTAATACTCACTATATGAAGACATTAGATAGTTTCCAATTTTCTCT
GGAACTACAAACAAGCTTGGAGATATTAAACTTTGAAACAGTATATCAATATTTTACTGCTTTTGAAATTAAAGCTTTGA
CTCCTTATAAAACTATTGGTACCCGAACAATCAAACTTATTCCAGCAATTATTGAGAAGAATAATCTAAAACTCTATTTA
GATCTTACAAATAATGACTTTATTTGCTCTTCTCTAGAATTTTCCATAAGAATTATTATTGGAAATCCATTGAAAAATGA
ATCATCCAATTCCTCAAATTCCAATCAAGAGATACTTTCAATTATGATTAATGATGAACGTAATGATACATCAGCTCTTA
GTAAACTTTGTCCATACAAGATCCACCCCAAGATTGCCGATTTCTATCCAAAAGAAGAAATCTCATATAAATTTAAAGAA
CCAAGAAAACTACCTGGACCTCTTCTCCCTTATGGCTTCTCTATCCTAATTTTACTATCACTATTCAAAATTATTCCTAT
TTGGAATAAGTTGTCCAAGATTGACCAAGGATTCTTCACTTTCATAAATAACGTCCCTTTCATTAAAGTAGCAACCTTCA
TTTCTTTGGCAATTTCTCTTTTCATTATCCTTTGCTACTGGCATACACTCAACATCTTCCAGTTTATGTACATCTTCACC
CCAGCAATATGTATCTTTTTTGTTTTATTAAAACTCTCTCTTAGAAACTTTAATTATCCAAGCTTCGGCAATGAAAAGTC
CAAGTCTGACTAG

**Genomic Sequence**

2013 bp

ATGGGAAAAATACTTGTTTCTTTAATATTATTTGTTTTTGCCTTAATAAATAAGGTACAAACCCAATTCGTTAATATTTT
ACCAGAGGTACCGAGTGATATCACGCAAGAAATAGGTACATGTGATGCATTATTATTAAAATTACTAGCAACAAATTATG
TCCCAAAATCAGAAACTAACTCACTTTGTAACATTTTCAGAAATAAGATTCAAAGGATAGATACTCCATTAAATGATTTA
AATTGCTCAATATTTGGTTCGATACTAGTTCAATGTGATCTCAAATTAGCAATTACTAGTTCTGAAATCTTAAGACAATA
TTTTATGAAATCAACTAAGCCAACCTTAAAAGAGCTATATTACTTAACAACTATACTATCCAATTTACCACAGGAAATTA
GTTCAGTTTTAGGAAGAGACGATATAATAGAAGACTCAAATGAGATTATTAATGATAAGTTAAGTATTATCAAGAAACAA
ATTTCAAGTAATCAAATATTAGATTATGAAGAAATCTCATTTTTATTTGGATCAATTTCTCAGTTAATTCGATTAGGATC
TAGAAATGTACAACATCTTAAAATGATGAATACAATCTTGGAATTGCAAAATAAGAATCTCAAGTTTTATCCTATTGTAG
ACTTAGATACCTTTGTCCTCAACTTCTCACTTTTATCTAAAGCGTACTTCTTCTTAAGAAGACCCGTTGATTTTGAAGAT
ATCCCACTTTCCTTACAGCAAATGGCCTATAGTATTTCAAGAATTAAACTTAGAGATATAAAAAGTTTAACAGTTTTGAA
TGAATATATTGTCACTAGAAGACTGTTTATTGATTCTGGTATTTATTTGATACAAGGTTTCCAAAATACTAATCCCAAAA
ATATTGATGAGAATTTAGCTACAGTTATATTTTGTAGATTAGATGGATCACCAATTGAAAATTTATCTTTGGATTTATCA
AATACTAATACTTTTAAACTTGTAGGTATACCAAATACTTGCGAGTATCAACTTTTTTATAATATTGCAAATCAAACCGA
ACCATTTAATCTTGAGAAAGTGAAGAATGATTTGTCTAAATCTATTCCAAAGCTAAAAATCTCTAGTAGATCAGTCCTTG
CTAGAGAAGCCAATATTCCAATTCTTAATAAGAAAAGCTCCAATTTTGAGATCACAAAAGCCACTATTTACTTTGGTGAG
AAGTTAACATTGGGAAAATCTCAAGTTAAGGATATTAATACTCACTATATGAAGACATTAGATAGTTTCCAATTTTCTCT
GGAACTACAAACAAGCTTGGAGATATTAAACTTTGAAACAGTATATCAATATTTTACTGCTTTTGAAATTAAAGCTTTGA
CTCCTTATAAAACTATTGGTACCCGAACAATCAAACTTATTCCAGCAATTATTGAGAAGAATAATCTAAAACTCTATTTA
GATCTTACAAATAATGACTTTATTTGCTCTTCTCTAGAATTTTCCATAAGAATTATTATTGGAAATCCATTGAAAAATGA
ATCATCCAATTCCTCAAATTCCAATCAAGAGATACTTTCAATTATGATTAATGATGAACGTAATGATACATCAGCTCTTA
GTAAACTTTGTCCATACAAGATCCACCCCAAGATTGCCGATTTCTATCCAAAAGAAGAAATCTCATATAAATTTAAAGAA
CCAAGAAAACTACCTGGACCTCTTCTCCCTTATGGCTTCTCTATCCTAATTTTACTATCACTATTCAAAATTATTCCTAT
TTGGAATAAGTTGTCCAAGATTGACCAAGGATTCTTCACTTTCATAAATAACGTCCCTTTCATTAAAGTAGCAACCTTCA
TTTCTTTGGCAATTTCTCTTTTCATTATCCTTTGCTACTGGCATACACTCAACATCTTCCAGTTTATGTACATCTTCACC
CCAGCAATATGTATCTTTTTTGTTTTATTAAAACTCTCTCTTAGAAACTTTAATTATCCAAGCTTCGGCAATGAAAAGTC
CAAGTCTGACTAG

**Swp1 sequences** – T. vaginalis strain G3 (obtained from TrichDB)

**Transcript id:** rna_TVAG_476400-1

**Transcript length:** 826

**Protein length:** 259

**Genomic length:** 826

**Predicted Protein Sequence**

259 aa

MLALTLAAFVRAAEISAIKFIKYDDNVENMTLDAKPNQVVSLDLTENQTLIAKLEGLKGEAKHAFYVLEQGTYSIVENLQ
SKGTYAAKLNPRALAGLYKHPGEYSLKVSITYEGEKPIMTEIAKINFIANGEVIDNFTDVEWDFQKPHEQPGAFLVFVFE
VASFVPIFILLVLLLINGCNFGYFPRNFFDAIFSITFVVAFGGFLYYFIYFWKHIHFEEMLKQLCVIFPALLILLRLALI
GRAKMVARDVPAEEKVKTE

**Predicted RNA/mRNA Sequence (Introns spliced out; UTRs highlighted)**

826 bp utr

ATGCTCGCTCTTACCTTGGCCGCATTTGTCCGCGCAGCAGAGATAAGTGCAATTAAATTTATTAAATATGATGACAATGT
CGAGAATATGACTTTAGATGCTAAACCAAATCAAGTTGTTTCTTTAGATTTAACAGAAAATCAGACATTAATTGCCAAAC
TCGAAGGACTTAAAGGCGAAGCAAAGCATGCCTTTTATGTCCTTGAGCAAGGTACATACTCAATTGTTGAAAACCTTCAA
TCAAAAGGAACATATGCCGCCAAACTTAATCCACGCGCTTTAGCAGGTCTTTACAAACATCCAGGTGAGTATTCACTTAA
AGTTTCAATCACTTACGAAGGTGAAAAGCCAATTATGACAGAAATTGCTAAAATCAATTTCATCGCCAATGGCGAAGTTA
TTGACAACTTTACAGACGTCGAATGGGACTTCCAAAAGCCACATGAACAACCAGGAGCTTTCCTTGTCTTCGTATTTGAA
GTAGCTTCCTTTGTTCCAATCTTTATCTTGTTAGTTCTTCTTTTAATCAACGGATGCAACTTCGGATACTTCCCAAGAAA
CTTCTTCGATGCTATATTCTCAATTACTTTTGTAGTTGCTTTTGGCGGATTCTTATACTATTTCATTTATTTCTGGAAGC
ATATTCACTTCGAGGAGATGCTCAAGCAATTATGCGTAATATTCCCAGCTTTACTTATCTTACTCCGCCTTGCTCTTATT
GGCCGTGCTAAGATGGTTGCTAGAGATGTTCCAGCCGAGGAAAAGGTTAAAACAGAATAAatattgcattgatcgttctt
Aaaattactaaacattaaatctaatt

**Genomic Sequence (UTRs highlighted)**

826 bp utr

ATGCTCGCTCTTACCTTGGCCGCATTTGTCCGCGCAGCAGAGATAAGTGCAATTAAATTTATTAAATATGATGACAATGT
CGAGAATATGACTTTAGATGCTAAACCAAATCAAGTTGTTTCTTTAGATTTAACAGAAAATCAGACATTAATTGCCAAAC
TCGAAGGACTTAAAGGCGAAGCAAAGCATGCCTTTTATGTCCTTGAGCAAGGTACATACTCAATTGTTGAAAACCTTCAA
TCAAAAGGAACATATGCCGCCAAACTTAATCCACGCGCTTTAGCAGGTCTTTACAAACATCCAGGTGAGTATTCACTTAA
AGTTTCAATCACTTACGAAGGTGAAAAGCCAATTATGACAGAAATTGCTAAAATCAATTTCATCGCCAATGGCGAAGTTA
TTGACAACTTTACAGACGTCGAATGGGACTTCCAAAAGCCACATGAACAACCAGGAGCTTTCCTTGTCTTCGTATTTGAA
GTAGCTTCCTTTGTTCCAATCTTTATCTTGTTAGTTCTTCTTTTAATCAACGGATGCAACTTCGGATACTTCCCAAGAAA
CTTCTTCGATGCTATATTCTCAATTACTTTTGTAGTTGCTTTTGGCGGATTCTTATACTATTTCATTTATTTCTGGAAGC
ATATTCACTTCGAGGAGATGCTCAAGCAATTATGCGTAATATTCCCAGCTTTACTTATCTTACTCCGCCTTGCTCTTATT
GGCCGTGCTAAGATGGTTGCTAGAGATGTTCCAGCCGAGGAAAAGGTTAAAACAGAATAAatattgcattgatcgttctt
Aaaattactaaacattaaatctaatt

**OST3/6 sequences** – T. vaginalis strain G3 (obtained from TrichDB)

**Transcript id:** rna_TVAG_374690-1

**Transcript length:** 548

**Protein length:** 159

**Genomic length:** 548

**Predicted Protein Sequence**

159 aa

MGKSFKDIILAPFAEDRFFVPELEIPISTIKIPSMTICMLIVFSSFMVISAGTIFCWVHNSPFIGGQYDQNGKIRTLVFS
EGMSWQFGAEGFLASMVYVMTAFSFLASYYVFKHQNDNPNDPTIIAAKVFGYTSPVWIILMIMTFRSKLRQYFPTPFPQ

**Predicted RNA/mRNA Sequence (Introns spliced out; UTRs highlighted)**

548 bp utr

ATGGGAAAGAGCTTCAAAGATATCATACTTGCTCCATTTGCAGAAGACCGATTCTTTGTTCCAGAGTTAGAAATTCCAAT
TTCAACCATTAAAATCCCATCAATGACAATTTGCATGTTAATCGTATTCTCATCTTTCATGGTCATTTCAGCCGGTACAA
TTTTCTGCTGGGTCCATAACTCACCATTTATTGGTGGTCAATATGATCAAAATGGCAAGATCAGAACCCTCGTTTTTTCT
GAAGGTATGTCTTGGCAATTTGGTGCTGAAGGTTTTCTGGCAAGCATGGTATACGTTATGACAGCATTTTCCTTCTTAGC
ATCATATTACGTTTTCAAGCATCAAAATGATAATCCAAACGATCCAACAATTATTGCGGCAAAAGTTTTTGGATATACTT
CTCCTGTTTGGATCATTTTAATGATTATGACATTCAGAAGTAAACTCCGCCAGTATTTCCCAACGCCATTCCCACAATAA
tcatttcacttctttatttattacgatacaatgataattgtatacatattcgtatttttttaatgttgg

**Genomic Sequence (UTRs highlighted)**

548 bp utr

ATGGGAAAGAGCTTCAAAGATATCATACTTGCTCCATTTGCAGAAGACCGATTCTTTGTTCCAGAGTTAGAAATTCCAAT
TTCAACCATTAAAATCCCATCAATGACAATTTGCATGTTAATCGTATTCTCATCTTTCATGGTCATTTCAGCCGGTACAA
TTTTCTGCTGGGTCCATAACTCACCATTTATTGGTGGTCAATATGATCAAAATGGCAAGATCAGAACCCTCGTTTTTTCT
GAAGGTATGTCTTGGCAATTTGGTGCTGAAGGTTTTCTGGCAAGCATGGTATACGTTATGACAGCATTTTCCTTCTTAGC
ATCATATTACGTTTTCAAGCATCAAAATGATAATCCAAACGATCCAACAATTATTGCGGCAAAAGTTTTTGGATATACTT
CTCCTGTTTGGATCATTTTAATGATTATGACATTCAGAAGTAAACTCCGCCAGTATTTCCCAACGCCATTCCCACAATAA
tcatttcacttctttatttattacgatacaatgataattgtatacatattcgtatttttttaatgttgg

**OST3/6 sequences** – Entamoeba histolytica strain HM-1:IMSS (obtained from AmoebaDB)

**Transcript id:** EHI_096150A

**Transcript length:** 378

**Protein length:** 125

**Genomic length:** 378

**Predicted Protein Sequence**

125 aa

MASTITRFFQKNIFSFILVSYAIIMAGIFYDIIIEPPGTGSVIDKYGNIKPETIMKGRHNGQYVVEGICASIFFVMIAGG
MVIVDKSISMTEADRKKPLFAVGGVVATSFGLLMIYFFAKTKFGF

**Predicted RNA/mRNA Sequence (Introns spliced out)**

378 bp

ATGGCTTCAACAATAACAAGATTCTTTCAAAAGAATATTTTTTCATTCATATTAGTTTCTTATGCCATTATTATGGCAGG
AATATTCTATGATATTATTATTGAACCTCCTGGAACAGGAAGTGTTATTGATAAATATGGAAATATTAAACCTGAAACTA
TCATGAAAGGAAGACACAACGGGCAATATGTTGTTGAGGGAATCTGTGCATCTATATTTTTCGTTATGATTGCTGGTGGA
ATGGTTATAGTTGATAAGTCAATATCAATGACAGAGGCTGATAGAAAAAAGCCTTTATTTGCAGTGGGTGGTGTAGTTGC
TACATCTTTTGGATTGTTAATGATTTATTTCTTTGCAAAAACAAAATTTGGATTTTAA

**Genomic Sequence**

378 bp

ATGGCTTCAACAATAACAAGATTCTTTCAAAAGAATATTTTTTCATTCATATTAGTTTCTTATGCCATTATTATGGCAGG
AATATTCTATGATATTATTATTGAACCTCCTGGAACAGGAAGTGTTATTGATAAATATGGAAATATTAAACCTGAAACTA
TCATGAAAGGAAGACACAACGGGCAATATGTTGTTGAGGGAATCTGTGCATCTATATTTTTCGTTATGATTGCTGGTGGA
ATGGTTATAGTTGATAAGTCAATATCAATGACAGAGGCTGATAGAAAAAAGCCTTTATTTGCAGTGGGTGGTGTAGTTGC
TACATCTTTTGGATTGTTAATGATTTATTTCTTTGCAAAAACAAAATTTGGATTTTAA

## Supplementary figures

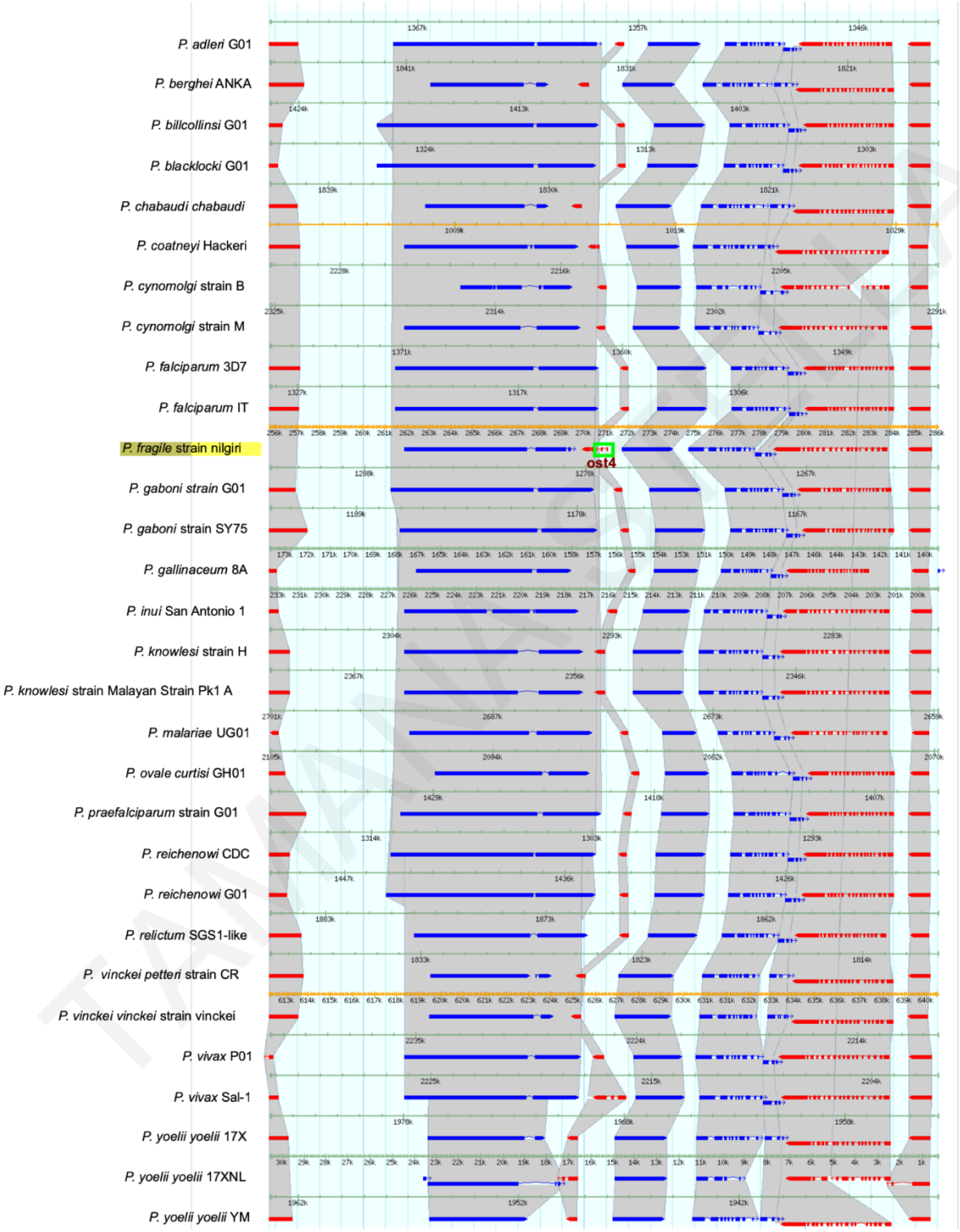



**Figure 52**: Syntenic region surrounding gene AK88_01616, encoding a putative OST4 subunit in P. fragile strain nilgiri. The green box indicates the putative ost4 gene lying within a region of conserved orthologs among all Plasmodium genomes.

# Appendix IV

**Dissecting sequence and structural features of compositionally biased regions in the Protein Data Bank**

- **Datasets**
  - PISCES
  - PDB
  - DSSP
  - NACCESS
  - CAST
- **Results**
  - Excel files
  - Clustering
  - Analysis
    - Sequence features
    - Structural features
  - Fischer test
- **Source code**
  - DSSP
  - NACCESS
  - CBRs signatures
  - Statistics