Title

Genomic approaches to investigate the control of gene expression in the human malaria parasite, Plasmodium falciparum

Author

Ahyong, Vida Lou

Publication Date

2015

Genomic approaches to investigate the control of gene expression in the human malaria parasite, Plasmodium falciparum

by

Vida Lou Ahyong

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Genetics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedicated to Troy Morris and my Family

Acknowledgements

I would like to thank my advisor, Joe DeRisi for his support, mentorship, and creating such a fun and exciting place to learn. His enthusiasm for science is infectious and I will forever be grateful that he allowed me to join the lab and explore such fascinating and important biological questions. I promise to 'hang in there' and not to '**** it up'. Thank you from the bottom of my heart. You are truly an inspiration.

I am so very grateful to all my science mentors and supporters. First, I'd like to acknowledge Aimée Dudley, Catherine Ludlow, and Cecilia Garmendia for training me during my first research experience at the ISB. Thank you for encouraging me to continue on with graduate school and being wonderful role models. I'd like to thank my classmates, the incoming Tetrad class of '09! I am so thankful for your support in such a difficult first year and for being wonderful friends and colleagues throughout our graduate career. I can't wait to see where all of your crazy smart people will end up! I would like to thank the members of my thesis committee, Jonathan Weissman, Christine Guthrie, and Alan Frankel for all the guidance, encouragement, and helpful suggestions.

I would like to especially thank all the really wonderful people from the DeRisi and Weissman labs that I feel privileged to call my friends. The Gaggle: Flor Caro, Gloria Brar, and Emily Wilson, you are the epitome of hard working, selfless, brilliant, fun, and beautiful women and I am so happy that we are friends. To the amazing DeRisi Postdocs: Polly Fordyce, Ellen Yeh, Charlie Kim, Mark Stenglein, and J. Graham Ruby; your unbelievable mentorship has been an invaluable asset to my training, thank you for all your support and putting everything in perspective for me when times were tough. To my computational quad buddies: Miguel Betegon, J. Graham Ruby, Sharon Chao-

Wooten, Michelle Dimon, Peter Skewes-Cox, Greg Fedewa, Brian O’Donovan, Emily Crawford, Jenn Mann, Chris Sheridan and Bullet (the best lab dog), thank you for all the laughs, commiseration, youtube video breaks, printer messages, mantid rearing, food club and whiskey sharing. You make the day-to-day so much more enjoyable. Thank you to our lab managers: Jenn Mann, Manny DeVera, and Tara Christiansen for making sure things ran smoothly in lab and always with a smile. And to all the others that have made the lab a better place with your wisdom, hard work, and friendship: Alex Greninger, Danny Ebert, Nathan Yozwiak, Brian Baxter, Calvin Jan, Martin Kampmann, Manuel Leonetti, and Stirling Churchman.

I am eternally grateful to my family: Virginia Ahyong, Eduardo Ahyong, and Edward Ahyong. We have been through so much together and I want to thank you for continually supporting me despite my stubbornness and mood swings!

Finally, I want to thank my most amazing and wonderful life partner and future husband, Troy Morris. We met at the beginning of my graduate career and despite all my late nights, panic attacks, and stress of grad school, you always have supported me and still you shower me with unconditional love every single day. Thank you for always laughing at my bad jokes, picking up on my hints, and always following your passion.

Genomic approaches to investigate the control of gene expression in the human malaria parasite, *Plasmodium falciparum*

by:

Vida Lou Ahyong

Abstract

The recent advances in genomic sequencing technologies in the past decade have enabled the unprecedented ability to investigate infectious diseases and organisms that were traditionally difficult to study in the laboratory. In particular, the human malaria parasite, *Plasmodium falciparum,* is a unique challenge because of the difficulty to grow these obligate intracellular parasites during the complex blood stages. Yet despite these challenges, genomic approaches have allowed us to investigate the parasite in an unbiased and comprehensive manner. The work described here uses DNA sequencing, RNA sequencing, high-throughput screening, and bioinformatics tools to understand the fundamental biological processes of the malaria parasite such as mechanisms of drug resistance, regulation of translation, and determinants of translational efficiency. A mechanistic understanding of how the parasite can escape antimalarial drug pressure is valuable for future drug development and the design of new drugs that avoid known parasite resistance mechanisms. To this end, we have developed a bioinformatics pipeline that will identify mutations such as copy number variations or single nucleotide polymorphisms that confer resistance to various drugs *in vitro*. The process of translation in malaria is poorly understood as the technologies for measuring translation or protein products are limited to measuring only the most abundant proteins. To address this problem, we took a genome-wide approach of ribosome profiling which quantitatively measures transcription and translation simultaneously for the entire genome. This study led to many interesting questions revolving around how the *P. falciparum* ribosome functions differently from other eukaryotic ribosomes and what *cis*-acting sequences determine the efficiency of translation. To answer these questions, we developed an *in vitro* translation system derived from cell free lysates of *P. falciparum* to both screen for drug compounds that specifically inhibit malaria translation and not other eukaryotic translation and we have employed this assay to find specific sequences found primarily on the 5' untranslated region of an mRNA that can modulate translation. In total, this work addresses the regulation of gene expression in the asexual blood stages of this medically relevant parasite.

Table of Contents

List of Tables

Chapter 2

Chapter 3:

Chapter 5:

Table of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 1: Introduction

Malaria continues to take a devastating toll on global health with approximately 3.3 billion people at risk of infection and disease and 1.2 billion at high risk (>1 in 1000 infections per year)[1]. In 2013 an estimated 198 million cases occurred leading to over 500,000 deaths for that year. Efforts to eradicate malaria are hindered by several complex factors one of which is the emerging threat of drug resistance to the current roster of antimalarials. One of the most pertinent examples is the evidence of resistance to the powerful antimalarial, Artemisinin, in the Thai-Cambodian border[2][3][4]. To combat this threat, many research efforts have focused on the discovery of new compounds with antimalarial properties. Drug discovery can be achieved through either a top-down or bottom-up approach. In the top-down approach, large collections of compounds can be screened for antimalarial activity by growing parasites and assaying how well the parasite can grow in the presence of the drug. Our lab has performed these types of screens by assaying for parasite growth after a 72 hour incubation in drug which have resulted in lead compounds such as Propafenone, identified as a potent antimalarial from a screen against ~2000 compounds in the MicroSource collection[5]. The advantage to this approach is the ability to screen collections indiscriminately with the hopes of finding many lead compounds and the disadvantage is the possibility that many of these compounds are targeting the same gene, ultimately limiting the number of lead compounds with diverse antimalarial effects. In the bottom-up approach, assays are designed to test the activity of a compound with a known target or pathway. A few examples in *P. falciparum* include assaying for the ability to form sexual stage precursor parasites called gametocytes or assaying for the loss of an essential organelle called the apicoplast[6][7]. The advantage for a bottom-up approach to drug screening is the pre-existing understanding of the mechanism of action for the drug and the avoidance of drug targets with known global resistance profiles. The greatest disadvantage for a bottom-up approach is in the selection of a novel assay that will be specific and potent against the organism of interest (*P. falciparum*) and not the human host. In order to achieve these requirements of specificity and potency, a substantial amount of familiarity with the basic biology of the organism is required.

Plasmodium falciparum biology and life cycle:

Human malaria is transmitted by one of five known protozoan *Plasmodium* parasites; *P. falciparum, P. vivax, P. ovale, P. malariae,* and *P. knowlesi*, with the most frequent being *P. vivax* and the most lethal being *P. falciparum*. The life cycle of the parasite is divided into two main organisms, the human host or the *Anopheles* mosquito vector with the asexual stages occurring in the human host and the sexual stages in the mosquito.

The life cycle begins with the bite from an infected female mosquito whose saliva carries hundreds of motile parasites called sporozoites, which migrate through the blood or the lymphatic system into the liver, to invade hepatocytes. Here, the parasites differentiate from a sporozoite into a dormant stage called a hypnozoite. During this period, thousands of parasites called merozoites can emerge from a single hynozoite to re-enter the blood stream. These released merozoites can now enter into their intraerythrocytic developmental cycle (IDC). This cycle is characterized by the infection of a single merozoite into an uninfected red blood cell where it develops through four general morphological stages; rings, early trophozoites, late trophozoites, and schizonts, before releasing 12-24 new merozoites from a single infected red blood cell to continue another round of infection. The IDC occurs in a 48 h period that results in the characteristic clinical features of high fever and chills when the synchronous populations of parasites are released from the infected blood cell and into the blood stream. Though this cycle can continue indefinitely until the person is treated with antimalarials or succumbs to the infection. To escape from the host and enter into its sexual stages of development, the parasite must differentiate from the blood forms into a terminal sexual stage precursor called a gametocyte. Though the many mechanisms of gametocyte induction and development have not been entirely revealed, there have been many studies that suggest that this process, called gametocytogenesis, is a general stress response and as such, the current laboratory protocols call for 'stressing' the cultures by limited feeding or growth at high parasitemia[8]. Once a parasite is terminally committed to gametocytogenesis, this parasite will undergo dramatic morphological changes in process lasting approximately two weeks [9]. The end result is a mixture of male and female mature gametocytes that are circulating in the bloodstream to be ingested by a blood meal from a female *Anopheles* mosquito. Once inside the mosquito, the temperature differential (though pH and xanthurenic acid are also involved) between the human host (37°C) and the mosquito (25°C) is sufficient for the male and female gametocytes to differentiate further into gametes which can now fuse for fertilization to form a diploid zygote which then develops to an ookinete where genetic recombination occurs during meiosis[10][11]. The ookinete then grows into an oocyst for 1-3 weeks in the midgut of the mosquito before undergoing multiple nuclear divisions to form hundreds of haploid sporozoites. These sporozoites travel to the salivary glands where they can now be transmitted back into a human host after a mosquito bite.

Of Genes and Genomes

Beyond understanding the unique life cycle of the parasite, it is critical to understand the mechanistic systems underlying the large morphological developmental processes to better inform future therapeutic interventions. For the last ~20 years, the DeRisi lab has been closely involved in understanding the complex cellular systems of the *P. falciparum* asexual blood stages. We have taken an approach that is unique in the parasitology field by not just studying one or two individual genes and their functions but we examine the entire genome and their pathways at once. By considering all genes in the *P. falciparum* genome at once we can take an unbiased look at which cellular systems or pathways are perturbed or modulated during drug exposure or during developmental changes.

The methods to analyze *P. falciparum* genome-wide data began with the publishing of the genome in 2002 which described a genome with ~5,300 genes, 14 nuclear chromosomes, 1 apicoplast chromosome, and 1 mitochondrial chromosome[12]. The nuclear chromosome contains a surprisingly biased nucleotide content of approximately 80% A-T richness for the 24 megabase genome which has proven to be a hurdle when considering the molecular and bioinformatics techniques applied to the study of this organism. In 2003, the DeRisi lab published the quintessential, high-resolution study of the entire *P. falciparum* transcriptome using microarrays, describing the continuous cascade of gene expression, where each gene is expressed just once during the 48 h life cycle in a just-in-time fashion[13]. This result along with additional microarray transcriptome studies of the IDC during drug or stress exposure seemed to suggest that transcription was not a major regulatory feature in the parasite[14][15]. Furthermore, the few proteomic studies performed afterwards suggested that a portion of the transcriptome had not been translated into protein, implicating post-transcriptional mechanisms as a major regulatory feature in the parasite, though the proteomic data was sparse in their final quantity of proteins measured[16][17][18].

Post-transcriptional gene regulation

Post-transcriptional gene regulation such as transcription abortion, retention or degradation in the nucleus, blocking of translation, and RNA degradation can be achieved through several mechanisms[19]. Some notable examples include RNA binding proteins that bind to a transcript and prevent ribosome initiation or elongation or *cis*-regulatory sequences that act as poor substrates for the ribosome. These mechanisms can be controlled on a local scale where just a single transcript is up- or down-regulated due to targeted regulation or on a global scale, where all translation and specifically the ribosome is inhibited. How do we determine if post-transcriptional regulatory mechanisms are at play in a system? Traditional molecular and biochemical techniques can be utilized to detect the presence of an mRNA such as Northern blotting or RT-qPCR to test for cases when mRNA could be subject to degradation[20]. In addition, the use of polysome profiling in which cell lysate is separated by density on a sucrose gradient can be useful when detecting whether a transcript is associated with high-density polyribosomes, a signal for robust translation, or low-density ribosomes, a signal for poor translation[21]. Yet these techniques suffer from the low throughput results because the researcher must design their assays to specifically target a known set of transcripts. The alternative is to use new genome-wide approaches that take unbiased measurements of transcription or translation. These approaches rely heavily on deep sequencing technologies such as Illumina short read sequencing which provides hundreds of millions of short read fragments from a single run that can be assembled or aligned to pre-existing genomes. In the DeRisi lab, we have specialized in using deep sequencing technologies to quantitatively address many outstanding questions in the understanding of the basic biology of malaria such as the mechanisms of antimalarial drug resistance, post-transcriptional gene regulation, and *cis*-acting regulatory features unique to the parasite. Chapter 2 and 3 will describe the deep sequencing approach and bioinformatic analysis I took to understand the genetic changes that confer drug resistance to two different classes of antimalarial drugs, a potent inhibitor of *Pf* dihydroorotate dehydrogenase, and the widely used artemisinins. Chapter 4 will describe a highly collaborative, multi-year effort to determine the extent of post-transcriptional regulatory mechanisms of the *P. falciparum* asexual blood stages. Chapter 5 will detail the development of an *in vitro P. falciparum* translation assay and how I used the assay to discover novel protein synthesis inhibitors. Finally, Chapter 6 will describe the preliminary work I have begun to understand the *cis*-regulatory features on mRNAs and how they influence translational efficiency.

References

1. World Health Organization, Global Malaria Programme, World Health Organization. World Malaria Report 2014. 2014.

2. Carrara VI, Zwang J, Ashley EA, Price RN, Stepniewska K, Barends M, Brockman A, Anderson T, McGready R, Phaiphun L, Proux S, Van Vugt M, Hutagalung R, Lwin KM, Phyo AP, Preechapornkul P, Imwong M, Pukrittayakamee S, Singhasivanon P, White NJ, Nosten F. Changes in the Treatment Responses to Artesunate-Mefloquine on the Northwestern Border of Thailand during 13 Years of Continuous Deployment. PLoS ONE. 2009 Feb 23;4(2):e4551.

3. Noedl H, Socheat D, Satimai W. Artemisinin-resistant malaria in Asia. N Engl J Med. 2009 Jul 30;361(5):540–541.

4. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, Tarning J, Lwin KM, Ariey F, Hanpithakpong W, Lee SJ, Ringwald P, Silamut K, Imwong M, Chotivanich K, Lim P, Herdman T, An SS, Yeung S, Singhasivanon P, Day NPJ, Lindegardh N, Socheat D, White NJ. <title> Artemisinin Resistance in *Plasmodium falciparum* Malaria </title>. N Engl J Med. 2009 Jul;361(5):455–467.

5. Weisman JL, Liou AP, Shelat AA, Cohen FE, Guy RK, DeRisi JL. Searching for new antimalarial therapeutics amongst known drugs. Chem Biol Drug Des. 2006 Jun;67(6):409–416.

6. Wu W, Herrera Z, Ebert D, Baska K, Cho SH, DeRisi JL, Yeh E. A chemical rescue screen identifies a Plasmodium falciparum apicoplast inhibitor targeting MEP isoprenoid precursor biosynthesis. Antimicrob Agents Chemother. 2015 Jan;59(1):356–364.

7. Sanders NG, Sullivan DJ, Mlambo G, Dimopoulos G, Tripathi AK. Gametocytocidal screen identifies novel chemical classes with Plasmodium falciparum transmission blocking activity. PLoS ONE. 2014;9(8):e105817.

8. Kooij TW, Matuschewski K. Triggers and tricks of Plasmodium sexual development. Curr Opin Microbiol. 2007 Dec;10(6):547–553.

9. Baker DA. Malaria gametocytogenesis. Mol Biochem Parasitol. 2010 Aug;172(2):57–65.

10. Dyer M, Day KP. Regulation of the rate of asexual growth and commitment to sexual development by diffusible factors from in vitro cultures of Plasmodium falciparum. Am J Trop Med Hyg. 2003 Apr;68(4):403–409.

11. Williams JL. Stimulation of Plasmodium falciparum gametocytogenesis by conditioned medium from parasite cultures. Am J Trop Med Hyg. 1999 Jan;60(1):7–13.

12. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome

sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002 Oct 3;419(6906):498–511.

13. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol. 2003 Oct;1(1):E5.

14. Llinás M, Bozdech Z, Wong ED, Adai AT, DeRisi JL. Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. Nucleic Acids Res. 2006;34(4):1166–1173.

15. Gunasekera AM, Patankar S, Schug J, Eisen G, Wirth DF. Drug-induced alterations in gene expression of the asexual blood forms of Plasmodium falciparum. Mol Microbiol. 2003 Nov;50(4):1229–1239.

16. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, Yates JR, Winzeler EA. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res. 2004 Nov;14(11):2308–2318.

17. Cui L, Lindner S, Miao J. Translational regulation during stage transitions in malaria parasites. Ann N Y Acad Sci. 2015 Apr;1342:1–9.

18. Zhang M, Joyce BR, Sullivan WJ, Nussenzweig V. Translational control in Plasmodium and toxoplasma parasites. Eukaryotic Cell. 2013 Feb;12(2):161–167.

19. Alberts B, editor. Molecular biology of the cell. 5th ed. New York: Garland Science; 2008.

20. Gerst JE, editor. RNA detection and visualization: methods and protocols. New York: Humana; 2011.

21. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci USA. 2003 Apr 1;100(7):3889–3894.

Chapter 2: Genomic characterization of Antimalarial Drug Resistance

This chapter is a reprint from the following reference:

Jennifer L. Guler*, Daniel L. Freeman* , Vida Ahyong, Rapatbhorn Patrapuvich, John White, Ramesh Gujjar, Margaret A. Phillips, Joseph DeRisi, Pradipsinh K. Rathod (2013) Asexual Populations of the Human Malaria Parasite, *Plasmodium falciparum*, use a two-step genomic strategy to acquire accurate, beneficial DNA amplifications. *PLoS Pathogens* 2013;9(5): e1003375. doi:10.1371/journal.ppat.1003375

* These authors contributed equally to this work

Author contributions:

Conceived and designed the experiments: JLG DLF JW MAP PKR. Performed the experiments: JLG DLF VA RP JW RG. Analyzed the data: JLG DLF VA RP. Contributed reagents/materials/analysis tools: PKR JD MAP. Wrote the paper: JLG DLF PKR.

Abstract

Malaria drug resistance contributes to up to a million annual deaths. Judicious deployment of new antimalarials and vaccines could benefit from an understanding of early molecular events that promote the evolution of parasites. Continuous in vitro challenge of *Plasmodium falciparum* parasites with a novel dihydroorotate dehydrogenase (DHODH) inhibitor reproducibly selected for resistant parasites. Genome-wide analysis of independently-derived resistant clones revealed a two-step strategy to evolutionary success. Some haploid blood-stage parasites first survive antimalarial pressure through fortuitous DNA duplications that always included the DHODH gene. Independently-selected parasites had different sized amplification units but they were always flanked by distant A/T tracks. Higher level amplification and resistance was attained using a second, more efficient and more accurate, mechanism for head-to-tail expansion of the founder unit. This second homology-based process could faithfully tune DNA copy numbers in either direction, always retaining the unique DNA amplification sequence from the original A/T-mediated duplication for that parasite line. Pseudo-polyploidy at relevant genomic loci sets the stage for gaining additional mutations at the locus of interest. Overall, we reveal a population-based genomic strategy for mutagenesis that operates in human stages of P. falciparum to efficiently yield resistance-causing genetic changes at the correct locus in a successful parasite. Importantly, these founding events arise with precision; no other new amplifications are seen in the resistant haploid blood stage parasite. This minimizes the need for meiotic genetic cleansing that can only occur in sexual stage development of the parasite in mosquitoes.

Author Summary

Malaria parasites kill up to a million people around the world every year. Emergence of resistance to drugs remains a key obstacle against elimination of malaria. In the laboratory, parasites can efficiently acquire resistance to experimental antimalarials by changing DNA at the target locus. This happens efficiently even for an antimalarial that the parasite has never encountered in a clinical setting. In this study, we formally demonstrate how parasites achieve this feat: first, individual parasites in a population of millions randomly amplify large regions of DNA between short sequence repeats of adenines (A) or thymines (T) that are peppered throughout the malaria parasite genome. The rare lucky parasite that amplifies DNA coding for the target of the antimalarial, along with dozens of its neighboring genes, gains an evolutionary advantage and survives. In a second step, to withstand increasing drug pressure and to achieve higher levels of resistance, each parasite line makes additional copies of this region. This second expansion does not rely on the random A/T-based DNA rearrangements but, instead, a more precise amplification mechanism that retains the unique signature of co-amplified genes created earlier in each parasite. Generation of multiple copies of the target genes in the parasite genome may be the beginning of other beneficial changes for the parasite, including the future acquisition of mutations.

Introduction

The emergence of chloroquine and Fansidar resistance contributed to resurgence of malaria in the 1970s and 1980s [1], [2]. Today, from an estimated 2 billion global clinical cases, ~0.5 to 1 million individuals die of malaria every year [3], [4], [5]. There is a growing concern that decreased effectiveness of artemisinin combination therapies in Southeast Asia will once again lead to even higher morbidity and mortality [6], [7], [8], [9], [10]. While point mutations and

DNA copy number variations have been associated with resistance to previously effective antimalarials [11], [12], [13], [14], [15], a detailed understanding of how haploid blood stages of malaria parasites acquire resistance to truly new antimalarials is critical for the effective management of this global disease.

Similar to what has been observed in clinical settings, Plasmodium falciparum malaria parasites are able to acquire resistance under controlled laboratory conditions [16], [17], [18], [19], [20], [21], [22], [23], [24]. Although parasites exposed to potent antimalarials do not show protective, real-time transcriptional responses [25], the targets of novel antimalarials are often definitively revealed in in vitro selected resistant parasites through novel mutations or copy number variations in the parasite genome [20], [21], [22], [24], [26], [27], [28]. Such selections are now routinely used to identify target pathways of new antimalarials, but early molecular steps leading to beneficial mutations remain unknown. Here, we use in vitro selections to understand how haploid malaria parasite populations, under continual antimalarial pressure, correctly acquire protective changes in their genome. These controlled laboratory selections with asexual blood-stage P. falciparum allow step-wise mechanistic dissection of independently evolving parasite cell lines in ways that are not possible in field isolates or other model organisms.

Results

Resistance was achieved by challenging P. falciparum parasites with DSM1, a new potent and selective inhibitor of dihydroorotate dehydrogenase (DHODH) [29] (see structure in inset of Fig. 1). In the initial DSM1 challenge, populations of 107 parasites developed resistance to 0.3 μM DSM1 (Fig. 1, Table S1). Four independently-derived clones, exhibiting ~5-fold resistance, were selected for further investigation (round 1 clones were designated C, D, E, and F; Table S2). Pair-wise comparative genomic hybridizations of DNA from parent versus DSM1-resistant clones revealed a single ~2- to 3-fold amplification event on chromosome 6 in all four round 1 clones (Fig. 2A, Fig. S1). The amplicon units ranged in size from 34 to 95 kb, covering 9 to 23 genes (Fig. 2B, C). As discussed below, the variation in the size of the amplicon unit between independently-selected clones provided a molecular fingerprint of each evolving parasite line. All amplicons in each round 1 clone included the DHODH gene (Fig. 2C; gene 19, PlasmoDB gene ID PFF0160c [30], Fig. 2D). DHODH mRNA and protein levels were correspondingly increased (Fig. S2), and mutations were not detected in the gene itself (Fig. S3). Whole genome sequencing of the parent Dd2 clone and clone C (see genome coverage rates in Table S3) confirmed the de novo Whole acquisition of the DHODH amplicon and the absence of causal point mutations hidden within individual amplicon units (Fig. 3A; Table S4, Fig. S4). In addition, resistance-associated point mutations were not detected anywhere else in the genome (Table S5, Table 1).

To learn how DSM1-resistant parasite populations efficiently arrived at these unique beneficial amplicons, we mapped the junction regions of each independently-derived DSM1 resistant clone. Based on the boundaries initially identified by mid-density microarray analysis (Table S6), we sequenced the DNA between adjoining amplicon units assuming a head-to-tail orientation, and identified long homopolymeric stretches of adenine or thymine (A/T tracks) between the 3′ end of one unit and the 5′ end of the second (Fig. S5). These A/T tracks fall mostly in intergenic regions at the edges of the P. falciparum amplicons, with clones C and F sharing exactly the same unit end point (Fig. 2C). The 3′ junction of the remaining two clones D and E exist in two separate introns of PFF0185c (gene 24, Fig. 2C and Table S7). Of the 8 independent events studied here (2 junctions for each of the 4 independently derived clones analyzed), all displayed A/T tracks at the junctions (0 events occurred at a non-A/T tracks). Since homopolymeric tracks of >10 bp make up 5% of the genome [31], the probability that all of the 8 independent events would randomly end with an A/T track is 1 in 25 billion.

Investigations into the orientation of amplicon units (i.e. head-to-head or tail-to-tail orientation) as well as whether they were situated outside of chromosome 6 (the original DHODH locus) were expected to provide mechanistic insight into what pathways may be acting at these locations. In a quantitative approach that was not achievable in earlier studies (either by our group (Fig. S5) or others [32], [33]), we acquired paired-end reads from whole genome sequencing that aligned to the junction regions of clone C and D. Histograms of read coverage displayed junctions that were consistent with both microarray and targeted sequencing results discussed above (Fig. 3A). Computationally-isolated reads from the above analysis failed to reveal recombination of the DHODH loci with A/T stretches elsewhere in the genome since reads from all possible junctions aligned to only two genomic locations: (1) the region that represents the reference genome match on chromosome 6 (Fig. 3B, red and yellow arrows) or (2) the opposite end of the amplicon unit (Fig. 3B and C, blue or green arrows; Fig. S6). These data formally prove that the tandem head-to-tail arrangement is the predominant outcome of the initial duplication in DSM1 resistant clones (Fig. 3D).

Based on outcomes from round 1 clones, we hypothesize that the initial resistance-conferring duplication around the DHODH locus arises from an imprecise, even chaotic, process involving mitotic rearrangement between random A/T tracks that are sprinkled at a high frequency across the genome. Importantly, there appears to be a second non-A/T based step for expanding P. falciparum amplicon numbers. When a DSM1 resistant parasite carried more than two units in a freshly-generated amplicon, each unit had the same length, genetic content, and junction regions. Conservation of these units in each independently-selected parasite clone suggested that, after an initial fortuitous duplication between random A/T tracks surrounding the DHODH locus, subsequent expansion of the founder amplicon involves precise homologous recombination that overrides chaotic, possibly unproductive A/T track-based mechanisms. This hypothesis was further tested by exposing round 1 clones to higher DSM1 concentrations (3 μM or 10 μM DSM1 in round 2 compared to 0.3 μM in round 1; Table S8). The resulting independent round 2 clones derived from clones C and D were ~15- to ~150-fold more resistant to DSM1 compared to the parent Dd2 (Fig. 4A, Table S9). Comparative genomic hybridizations also showed an increase of the founder DHODH amplicon in these round 2 clones (Fig. 4B and C, Table S6). Whole genome sequencing studies of the amplicon unit junctions of round 2 clones (see genome coverage rates in Table S3) again displayed solely the tandem head-to-tail orientation (Figs. 3B and C and S6). The precise maintenance of the respective founder amplicons in clones C and D is particularly remarkable given that resistance can be conferred by much smaller units as was observed in round 1 clones E and F (Fig. 2C).

To test whether the machinery that allows for faithful expansion of the DHODH amplicons would work with the same precision during deamplification, DSM1 resistant parasites were grown without antimalarial pressure over a long period of time. Overall, resistance of both round 1 and 2 clones initially decreased before stabilizing at ~2-to 3-fold (Fig. 4D–E, Tables S9 and S10). This observation suggested that there was a measureable fitness cost of maintaining higher levels of the DHODH amplicon, an idea that is consistent with other observations such as the normal growth rate of round 1 clones, the growth defect displayed by round 2 clones (Fig. S7A), and in some cases the increased growth rate following the removal of DSM1 pressure (Fig. S7B).

Similar to what has previously been observed with amplified loci on P. falciparum chromosomes 4, 5, and 11 [20], [34], [35], the step-wise decrease of DHODH copy numbers in the absence of DSM1 could be captured over time. Although the starting level of resistance differed, a gradual "dialing down" of the amplicon in the population to stable round 1 levels was observed for two independent C-derived round 2 clones (Fig. 4D–G). Furthermore, comparative genomic hybridization of clones isolated from these cultures grown in the absence of DSM1 for 3 months (DSM1 removal (DR) clones) showed that despite de-amplification, amplicon unit boundaries of C-derived clones were faithfully maintained (Fig. 4H and I, Table S6). Intriguingly, this implied that the pathway that relies on large stretches of homology to "dial up" the amplicon also controls the reverse action and does not allow the A/T track-based mechanism to disrupt amplicon units that were initially evolutionarily successful.

Discussion

The DSM1-based selection system offers a precise and reproducible experimental path to understand early events in the evolution of malaria drug resistance, and possibly many other aspects of parasite evolution. Observations of resistance mechanisms against this antimalarial clearly demonstrate that parasites favor de novo target amplification to achieve DSM1 resistance and, more generally, that two distinct steps are employed to arrive at beneficial DNA amplifications. In the first step, founder amplicons of independently-selected parasites are established through costly, random duplication of DNA between distant A/T tracks (Fig. 5, Step

1). In the second step, a more precise amplification mechanism efficiently "tunes" the copy numbers of preexisting duplications as needed in response to drug pressure (Fig. 5, Step 2) while avoiding disruption of initial beneficial changes. These detailed insights into how DSM1 resistance is established in malaria parasites raise new important questions regarding the evolution of this organism as a haploid population in a human host. In this environment, parasites encounter host immunity as well as antimalarial drugs, the later often arising in intermittent and changing ways. We believe that the parasite employs unique evolutionary strategies to win these battles without extensive damage in the haploid genome, even before the parasite has a chance to mix with other isolates during the diploid state in the mosquito. These issues are addressed in the subsections below.

Evolving as a Haploid Genome

The present findings underscore the extraordinary capability of the parasite to evolve during a human infection as a haploid asexual population. In nature, during a single human infection, a few hundred parasites entering the liver expand successfully to become many billions in the face of both drug and immune pressure. Once established in the blood, the parasites can increase and decrease in waves even without a reinfection. In order to evolve during these expansions, haploid parasites must do so with minimal damage to their genome. Similarly to what was first proposed for bacteria [36], the initial random sampling of duplications in the malaria genome under selective pressure serves as an effective first step to locate and identify genetic targets for resistance and generates enough of a foothold for the haploid parasites to proliferate under lethal pressure. The randomness of the initial duplication step in this organism is evident in our detailed molecular characterization of independently-selected resistant parasites from round 1 selections.

In addition, these early events also capture the large size of amplicons that are initially sampled (Fig. 2). Assuming one duplicated region of approximately 50–100 kb per parasite, in principle, it is possible to cover the entire 23 Mb P. falciparum genome with a few hundred parasites. However, this is clearly not the whole story: the large parasite populations of roughly a million cells required for a successful DSM1 resistance event (Table S1) points to possibly extensive number of "trial duplications" that are non-productive or even lethal to the parasite. The success rates of about 1:10,000,000 from round 1 selections against this completely novel evolutionary challenge (DSM1) are similar to a previous semi-quantitative estimation of the initial amplification rate in this organism that were inferred from challenges with a traditional aminoquinoline class of antimalarials in clinical use [37].

The few parasites that can identify a productive locus by chance in the first step then rely on a second more efficient step to achieve evolutionarily more robust levels of resistance (Fig. 5, Step 2). Based on survival numbers from round 2 selections (Table S8), this second step appears at least 100-fold more efficient once pseudo-polyploids have been established around a high priority locus. This second process also allows continual fine-tuning of amplicon unit numbers based on the level of antimalarial pressure (Fig. 4). For a haploid blood-stage parasite, when necessary, pseudo-polyploidy could even allow for the safe introduction of point mutations within the amplified region before amplicon units decrease to single copies (Fig. 5, Long Term). Indeed, during laboratory selections, amplifications of the target gene often are observed alongside point mutations in the same gene ([22], [28], [35] and our unpublished observations).

Both during in vitro selection and in natural human infection, these productive genomic alterations must take place independent of meiosis. Meiosis is the stage of the life cycle where “textbook” chromosomal crossover mixes different genomes from coinfections to bring together beneficial new traits and to remove damaged DNA in the progeny. However, the sexual stages at which meiosis occurs are not available to the parasite until the transmission of gamete stages to the mosquito. Prior to this stage of the life cycle, how does the evolving haploid parasite avoid large collateral damage as it is under pressure to change in the human? Our detailed characterization of clones from carefully-controlled independent experiments reveals a powerful evolutionary strategy to make precise changes in its genome while expanding in the human, away from the mosquito. At its core, the strategy involves the creation of a single significant new genetic amplification in an individual parasite, even as the entire genome is being sampled by a large starting population. Through controlled laboratory experiments, we directly observed that the amplicon responsible for resistance was the only new amplicon in every individual successful DSM1 resistant parasite. By avoiding adventitious new amplicons elsewhere in the genome, collateral damage is minimized during a time when meiotic cleansings are not available to the parasite. This precise genetic modification is not without cost: every event is accompanied by millions of parasites that do not amplify a useful portion of the genome and do not survive. Whether the initial rearrangements are occurring continuously during the life of the parasites or only in response to stress is a question that remains to be answered.

Benefits of an AT-rich Genome

The first step in the generation of the DHODH amplicon was clearly mediated by stretches of polyA sequences or polyT sequences (Figs. 5 (Step 1) and S5). Previously, similar homopolymeric A/T tracks have also been identified at the borders of naturally-occurring P. falciparum amplicons on chromosome 5 [32], [33], [38], solidifying the relevance of the current laboratory based observations to naturally occurring genomic amplifications in this organism. The A/T-based strategy revealed by these data is uniquely matched with the high AT content of the P. falciparum genome, which averages 81% AT but can reach upwards of 90% in introns and intergenic regions [39]. Exactly such approaches are probably not utilized by other Plasmodium species that cause human malaria or by other protozoan parasites. Of note, the genomes of the haploid blood stage of P. vivax, the second most prevalent human malaria species worldwide [40], averages ~60% A/T content [41] and Leishmania species that are prone to drug resistance and gene amplifications average ~40% A/T content [42], [43].

This A/T-dependent approach likely applies to many successful evolutionary selections of different P. falciparum parasites; genomic amplifications have been observed during the characterization of both lab-adapted and field-derived parasites from various regions of the world [15], [33], [34], [35], [38]. In the previous studies, however, the exact mechanistic origin of the genomic rearrangements was often ambiguous. First, amplicons were generated in response to antimalarials in clinical use, and independent founder events could not be distinguished from later rearrangements. Second, in some cases, parasites were isolated from clinical infections and thus information on both the clinical drug pressures and the life history of the parasite leading to observed mutational patterns (including passage through a mosquito and recombination with other genotypes) were lost. In the present study, since the DHODH amplicons were selected entirely in the asexual blood stage of P. falciparum, we can definitively conclude that the A/T track-mediated step is important in the initial acquisition of a new amplification and not in changing or rearranging amplicons later in evolution.

Potential Recombination Mechanisms

In addition to showing a general strategy of how a population of parasites narrows in on a resistance-conferring DNA locus, data from the present study points to the importance of two distinct biochemical processes that must operate in each parasite for overall evolutionary success. During replication, A/T tracks are known to cause polymerase pausing due to the rigid bend of the DNA structure [44], [45]. Events that follow could include the creation of a double strand break and recognition by a DNA repair pathway. Alternatively, the rigidness of A/T tracks may prevent adequate histone interactions, leaving DNA open to proteins that may trigger recombination pathways [46]. Recombination pathways generally require large regions of homology to mediate strand invasion but shorter stretches of repetitive bases have also been implicated in the initiation of various mitotic DNA rearrangements [47], [48], [49]. Recent studies of E. coli under stress also implicate very short G-rich sequences in template switching between stalled replication forks that leads to the duplication of large genomic regions [50]. In addition to a microhomology-mediated recombination pathway that repairs DNA breaks, a similar replication-based mechanism has been implicated in the generation of complex genomic rearrangements in yeast and humans [49], [51], [52]. Both of these processes appear to get by with extremely small stretches of homology (<10 bp), which are significantly shorter than the A/T tracks observed at the borders of P. falciparum amplicons (~30 bp, Fig. S5) [32], [33], [38]. A/T tracks as large as 60 bp are estimated to make up ~5% of the parasite genome [31], [39], [53] and although these sequences may take part in template switching or microhomology-mediated recombination as in other organisms, their significant length could also be enough to trigger more canonical recombination pathways requiring as little as 50 bp of homology [54].

Amplification Verses Point Mutation

Our selection studies with DSM1 show a clear preference for pathways that generate DHODH amplifications even though point mutations have been shown to prevent DSM1 binding to a recombinant catalytically active version of DHODH [55]. Given that round 2 clones display a broad range of DSM1 resistance (Table S9), we had to be sure that hidden point mutations (either in the DHODH gene itself or at other locations in the genome) were not contributing to survival in the presence of high levels of DSM1. Based on the very deep coverage of our whole genome sequencing studies (Table S3, Fig. S4), we are confident that even low frequency mutations within large amplicons would be detected. A few additional observations suggest that resistance is truly due to DHODH amplification and there are no other undetected causal mutations in the DSM1 resistant genome: 1) parasites maintain sensitivity to a number of additional antimalarials (Table S11) indicating that they are not employing a pleotropic resistance mechanism such as drug efflux, and 2) EC50 against DSM1 and DHODH copy number decrease in a parallel fashion (Fig. 4), which emphasizes the contribution of the chromosome 6 amplicons to the resistance phenotype (as opposed to changes in other regions of the genome). Despite our confidence in the sequencing data, we cannot rule out that variations in amplicon sizes, and related physiological effects, contribute to the relationship between amplicon copy number and drug resistance.

Why are genome amplifications favored over the acquisition of point mutations in the DSM1 model? The ease with which one can find the correct locus that confers drug resistance and the lack of severe penalties for expanding copy numbers in the neighborhood of the DHODH gene may allow the gene amplification path to dominate. Additionally, the pharmacodynamics of drug exposure during selections could also play a role in favoring amplifications over mutations. Continual, unrelenting drug pressure demands an immediate sustained solution from the parasite population with little tolerance for wrong guesses. Although there is a measurable fitness cost of maintaining many amplicons in the absence of drug pressure (Fig. S7), parasites thrive following the increase in copy numbers of dozens of genes by an order of magnitude. Intuitively, intermittent cycling of increasing antimalarial levels, as is applied in many in vitro selection systems, may provide parasites with a chance to acquire mutations that confer high level resistance, and possibly even lose a relevant amplicon that had served its purpose in the early stages of resistance evolution. Beyond this, the nature of the drug, the nature of the target, and its location in the genome could all contribute to the optimum pathway to resistance since continual, uninterrupted application of some antimalarials during laboratory selections (as utilized in our selection scheme) has successfully generated point mutations in various target genes ([19], [23] and additional unpublished work).

Memory and Generality

The present laboratory-controlled studies show that, in the absence of drug pressure, malaria parasites lose extra copies of amplicons. However, as often seen with field-derived amplicons (Table S13 and [56]), malaria parasites do not always revert back to single copies of the target gene but instead retain a low number of amplicons in the absence of drug pressure (Fig. 4). This act has important implications for future survival: when a parasite population encounters drug pressure that it has successfully overcome before, the population is poised to rapidly re-amplify relevant amplicons quickly and efficiently without heavy collateral damage associated with A/T-based reshuffling between genes near the target.

While the evolution of malaria parasites is studied here in the context of drug resistance as a selection force, the versatile parasite-specific mechanisms that are used to achieve evolutionary success must help the parasites deal with a diverse set of challenges. In the forward direction, acquisition of appropriate beneficial amplifications could help parasites survive antimalarial drugs but also other potential challenges such as host immunity [57]. It may not be a coincidence that the liver stage expansion of an incoming parasite first allows a few hundred parasites to expand to about 100,000 to a million parasites before the population faces unexpected immune-reactions or unusual erythrocyte genotypes of the human patient. In addition to a gain of genetic material through asymmetric recombination, the reverse direction could also have public health relevance. For example, deletions of specific genes in changing parasite populations could render rapid diagnosis tests ineffective [58] thereby misguiding diagnosis-based chemotherapy campaigns.

Conclusion

The initial two-step evolutionary strategy of P. falciparum identified here, likely driven by two different molecular pathways with different biochemical preferences, assists the parasite in finding productive solutions to new and unexpected evolutionary challenges. The strategy is well suited for a parasite population to evolve with minimum collateral damage in surviving cells, it can act to anticipate and mount a rapid response to repeat threats, and it may offer universal advantages to parasite populations that need to withstand multiple threats beyond drug pressure.

Materials and Methods

Parasite Culture

For each experiment, erythrocytic stages of P. falciparum (previously cloned HB3 or Dd2) were freshly thawed from frozen stocks and maintained as previously described [59], [60], [61]. Briefly, parasites were grown in vitro at 37°C in solutions of 2 to 2.5% hematocrit (serotype A positive human erythrocytes) in RPMI 1640 (Invitrogen) medium containing 28 mM NaHCO3 and 25 mM HEPES, and supplemented with 20% human type A positive plasma in sterile, sealed flasks, flushed with 5% O2, 5% CO2, and 90% N2. Cultures were maintained with media changes 3 times each week and sub-cultured as necessary to maintain parasitemia below 5%.

Initial DSM1 Challenge

The highest concentration of DSM1 to which clonal Dd2 and HB3 parasites could develop resistance was determined empirically as previously described [17]. To ensure genetically pure populations, aliquots of 10 infected erythrocytes of each clonal parasite line were allowed to proliferate to about 108 infected erythrocytes. From these populations, 102–107 infected erythrocytes were challenged in flasks with 0.1–10 μM DSM1 (results from 107 are displayed in Table S1). Additionally, 10 infected erythrocytes were challenged with these same concentrations, to ensure that DSM1 was effective and lethal. To confirm that the parasites could proliferate normally under these experimental conditions, one flask of 10 infected erythrocytes did not receive DSM1. This experiment was performed in triplicate, using three independent biological samples of both Dd2 and Hb3 clones. Media was changed 3 times each week (receiving fresh DSM1 each time) and cultures were split 1:2 once a week to guarantee a continuous supply for fresh erythrocytes during the experiment. Parasite proliferation was monitored by Giemsa-stained thin smear blood samples taken at each media change. Selection flasks were cultured until parasites were observed proliferating or until 90 days, whichever occurred first.

Selection of DSM1 Resistant Parasites (Rounds 1 and 2)

Using limiting dilution, 102 to 107 (Dd2) or 107 (HB3) populations of genetically pure parasites (see above) were plated across 24 wells of a 96-well plate (each clone was selected in quadruplicate on a single plate). Additionally, a control plate, containing 10 infected erythrocytes per 24 wells, was set up for each clone. To ensure that DSM1 was effective and lethal, the upper half of the control plate was treated with 0.3 μM DSM1. To show that the parasites could proliferate normally under the test conditions, the lower half of the control plate received no DSM1. Plates were cultured (as described above) until parasites were observed proliferating or up to 120 days, whichever occurred first. As soon as parasites were observed (Round 1 results are displayed in Table S2), the well contents were transferred to a new 10 ml culture flask for expansion, sample storage and sub-cloning. During this expansion, DSM1 resistant parasites were kept under continuous 0.3 μM DSM1 pressure and freeze-thawing and culturing for >1month at a time was avoided as much as possible. Four DSM1 resistant clones isolated in round 1 were submitted to another round of selections (round 2). Parasite populations of 10 and 107 were selected with 1, 3.3 and 10 μM DSM1 as described for round 1. In addition, 105 parasites were also challenged with 3.3 μM (Round 2 results are displayed in Table S8). Resistant parasites were isolated as described above before sub-cloning for further analysis.

Parasite Sub-cloning

To isolate genetically pure populations of DSM1 resistant parasites for further analysis, aliquots of 10–20 infected erythrocytes were plated across an entire 96-well plate. These plates were maintained (as described above) and as soon as parasites were observed proliferating, the well contents were extracted from the plate and transferred to a new 10 ml culture flask for further expansion, sample storage and analysis.

EC50 Determination by Hypoxanthine Uptake Assay

A parasite solution at 0.5–1% parasitemia (0.5% hematocrit) from the clone of interest was plated into a 96-well culture plate. An appropriate range of concentrations of DSM1 (from 0.02–200 μM), depending on the level of resistance of the parasites being tested, were then added to the parasites (because of solubility issues, 100× DSM1 concentrations (in 100% DMSO) were first diluted 1 : 10 into RPMI (final 10% DMSO) before being diluted again into the parasite-containing wells (final 1% DMSO)). Each concentration of interest was performed in triplicate and included solvent-only controls. After incubating for ~48 hours, wells were pulsed with 0.35 μCi each of 3H-hypoxanthine. Following an additional 24–40 hours, well contents were extracted and radioactivity was measured. Parasite proliferation in each test well was expressed as a percentage of the solvent control well. EC50 values were fit using the GraphPad PRISM software, according to the equation: Y=Bottom+(Top-Bottom)/(1+10((LogEC50−X) * HillSlope)).

Genomic DNA Isolation for Downstream Genomics Methods

For microarrays and quantitative PCR (qPCR) protocols, clonal asynchronous P. falciparum-infected erythrocytes were lysed with 0.15% saponin (Akros) for 5 min and genomic DNA (gDNA) was extracted using the DNeasy kit (Qiagen) according to the manufacturer's instructions. For whole genome sequencing, clonal P. falciparum cultures (30 mls in T75 flasks, 3% hematocrit) were synchronized with 5% sorbitol for two consecutive cycles (~45 hrs apart) and then once more (3–4 hr later) before harvesting for gDNA purification. These highly synchronous cultures (~3% parasitemia at >90% rings) were washed with PBS and frozen at −80°C prior to red blood cell lysis with saponin as above. Isolated parasites were washed 3× with PBS before resuspension in 150 mM NaCl, 10 mM EDTA, and 50 mM Tris-HCl pH 7.5. Parasites were lysed with 0.1% L-loril sarkosil (Teknova) in the presence of 200 μg/ml proteinase K (Fermentas) overnight at 37°C. Nucleic acids were then extracted with phenol/chloroform/isoamyl alcohol (25:24:1) pH 7.8–8.1 (Acros) using phase lock tubes (5 Prime). Following RNA digestion (with 100 μg/ml RNAse A (Fermentas) for 1 hr at 37°C), gDNA was extracted twice more as above, once with chloroform, and then ethanol precipitated by standard methods.

DNA Microarrays and Comparative Genomic Hybridization (CGH)

Spotted DNA microarrays (used for both CGH and expression analysis) consisted of 10,416 −70mer oligonucleotides designed from the P. falciparum 3D7 sequence with increased coverage for long ORFs [62]. Additional custom oligonucleotides were included in the microarray to increase coverage of genes involved in folate and nucleic acid metabolism. DNA was spotted on poly-lysine coated slides and post-processed using methods described previously [25], [63]. For hybridizations on spotted DNA microarrays, 5 μg of gDNA from each clone was sheared, labeled with 5-(3-aminoallyl)-2′ -deoxyuridine-5′ -triphosphate, and coupled to Cy-dyes as was done previously [64]. Uncoupled Cy-dyes were removed using the DNA Clean and Concentrate-5 kit (Zymo Research) before hybridization to the microarray at 62°C for 16–18 h. After washing, slides were dried, scanned at 10 μM resolution using the GenePix 4000B scanner and fluorescent images were quantified with GenePix Pro 3.0 (Axon Instruments). Further analysis, including normalization and statistical methods were performed as described previously [25]. Spotted microarray data are presented in MIAME-compliant format on the NCBI-based Gene Expression Omnibus (GEO) database (Accession # GSE35732).

Commercially manufactured mid-density CGH microarrays containing 385,585 oligonucleotide probes ranging in size from 15- to 45-mer were purchased from NimbleGen Systems, Inc. These microarrays are sufficient to detect copy number variations but not single nucleotide polymorphisms (SNPs) [65], [66]. For hybridizations to mid-density microarrays, gDNA was labeled with Cy3 and Cy5-labeled random nanomers (Trilink Biotechnologies) and hybridized to the current CGH design Plasmodium_3D7_WG_CGH as described previously [65] except hybridization was performed overnight (~16–18 h) in a 42°C water bath and microarrays were dried and scanned as above (at 5 μM resolution). Normalization and analysis was performed using NimbleScan version 2.6 (SegMNT CGH) and plotted using GraphPad PRISM. Mid-density microarray data are presented in MIAME-compliant format on the GEO database (Accession # GSE37306).

Quantitative PCR

For DHODH qPCR, two separate sets of primers were used to amplify a 206 bp amplicon beginning at nucleotide +656 of the DHODH coding sequence (DHODH front), and the second set amplified a 158 bp amplicon beginning at nucleotide +1423 of the DHODH coding sequence (DHODH rear) (see Table S11 for all primer sequences). The qPCR protocol was 95°C for 10 min, followed by 39 rounds of 95°C for 15 sec and 60°C for 1 min. For all experiments, we performed melt curves (55°C to 85°C in 0.5°C steps with 1 s hold at each step) to ensure a single amplicon was produced, and standard curves (10× dilution ladders of Dd2 gDNA) to determine the amplification efficiency. Relative copy number was determined for 1 ng of gDNA, using the Pfaffl method [67] according to the equation (Etarget)ΔCt, target (control−test)/(Eref)ΔCt, reference (control−test), where Seryl t-RNA Synthetase (PF07_0073) and 18 s Ribosomal RNA (MAL13P1.435) served as reference genes. DSM1 resistant clones served as the test, and the Dd2 parent served as the control. Significance was determined from multiple experiments with one-way ANOVA analysis and values from individual clones were compared using the Tukey's Multiple Comparison Test in GraphPad PRISM.

Whole Genome Sequencing (WGS)

I. Library Preparation

Illumina-compatible paired-end libraries were prepared from 50 ng gDNA (see isolation methods described above) using the Nextera DNA Sample Prep Kit (Epicenter Biotechnologies) according to the manufacturer's instructions except that we restricted the bridge PCR step to 6 cycles (instead of 9) and modified the extension step to 65°C for 6 min. Illumina-compatible adapters containing unique barcodes were used at this step instead of Nextera Adaptor 2 so that multiple samples could be run in the same lane of a flow cell by index read sequencing (IDX1=‘CGTGAT’: D73-1, IDX2=‘ACATCG’: Clone C, IDX3=‘GCCTAA’: Dd2, IDX4=‘TGGTCA’: C710-1b, IDX5=‘CACTGT’:C710-2a). Library fragments from 360 to 540 bp were then size selected on a 5 XT DNA 750 chip using the Lab Chip XT system (Caliper Life Sciences). A final limited-cycle PCR step (Klentaq LA DNA Polymerase (Sigma-Aldrich) with 80% A/T dNTPs) was performed with the outer sequencing adapters (6 cycles of 95°C for 10 sec, 58°C for 30 sec, 60°C for 6 min) in order to enrich for sequence-ready fragments. Prior to cluster generation, library concentrations were confirmed using a high sensitivity DNA Bioanalyzer (Agilent) and qPCR (with Nextera adapter sequences) and samples were pooled at 2 nM in sets of 3. Cluster generation was performed using the cBot HiSeq Cluster Kit v2 (Illumina, Inc.) at a final concentration of 6–8 pM and density of >400 k/mm2. Resulting flow cells were run using a v2 HiSeq flow cell on the HiSeq 2000 (Illumina, Inc., ~90 million reads per lane,Genbank Accession #SRA052245.2).

II. Basic analysis

Sequencing reads from individual libraries were separated according to their unique barcodes (introduced during library generation). All reads were aligned to the 3D7 reference genome (PlasmoDB v7.1) using Bowtie [68], allowing a single mismatch for unique reads only. Reads that aligned to multiple regions of the genome were discarded. Genome coverage was estimated as a percentage of the 3D7 genome that was covered by a certain number of reads (see Table S3 for coverage rates). Copy number variations (both amplifications and deletions) in round 1 and 2 clones were identified using histograms of normalized read coverage per million reads aligned over the genome using the Integrated Genome Browser (www.broadinstitute.org/igv). By examining histograms of read coverage across the genome, we detected two deletions evident on chromosomes 2 (position 61539–105810) and 9 (C clone, 1457193–1473789; C710-1b, 1379103–1474063; C710-2a, 1457258–1474460 and D73-1, 1393139–1473966) that were likely due to extended in vitro culture [69], [70] and were not considered further. In addition, two amplicons (chromosome 5 (position 888060–970425) and 12 (971307–976534)) that are well described in lab-selected and field-isolated clones (reviewed in [71]) were detected but not considered to contribute to the phenotype because their levels fluctuated between different resistant clones (Table S13).

SNPs were identified by calculating nucleotide frequency for every position in the 5 genomes sequenced independent of the reference genome nucleotide call. These frequencies were used to call the consensus nucleotide and specify amino acid changes if the position is in a coding region. Each clone was independently subjected to this analysis and ultimately compared to the sensitive Dd2 strain to make a ranked list of discordant SNPs. The top 100 SNPs per chromosome were filtered to identify non-synonomous SNPs in exons covered by >5 reads and present in >90% of reads (those from known hypervariable genes, such as pfEMP, rifin, var, and stevor were excluded). These lists were compared between resistant clones and the Dd2 sensitive clone in order to identify SNPs that could be contributing to DSM1 resistance (Table S5). Because the Dd2 sequenced during these studies was not the immediate parent of the DSM1 resistant clones, we verified five SNPs that were present in both round 1 and 2 clones. We performed PCR-directed sequencing (primers listed in Table S11) in multiple parasite clones including 3D7, two Dd2 clones (new, clone acquired from MR4 (MRA-156, MR4, ATCC Manassas Virginia) in June 2011; old, clone used in the lab for several years), all four round 1 clones, and several round 2 clones (Table 1). In addition, these SNPs were investigated in a clone from a recent independent DSM1 selection in which the parent Dd2 clone was known (Dd2 (new) and NS clone 1, Fig. S9). The optimized PCR protocol (95°C for 5 min, 30 rounds of 95°C for 30 sec, 55°C for 1.5 min, and 72°C for 2 min, with a final extension of 72°C for 10 min) gave a single amplified product of the expected size (DHODH-F/R, Table S12). Amplified product was sequenced using an additional 6 internal primers (Seq1–6. Table S12).

III. Detecting Mutations in Amplified Regions

To detect SNPs from whole genome sequencing reads in amplified regions, we performed a local BLAST (NCBI, version 2.2.25+) to align reads to nucleotide databases of the amplified region on chromosome 6 with a minimum alignment length of 50 nucleotides and an e-value of <10−3. Ungapped alignments were searched for SNPs by calculating nucleotide frequencies per position in the database. SNPs were filtered by percent frequency per nucleotide of the resistant clone over the parent Dd2 and we narrowed our focus on those within open reading frames with a minimum of 20 reads covering any suspected SNP position (presented in Table S4). In addition, Bowtie alignments were converted to the BAM file format for viewing in Integrated Genome Browser and aligned to the 3D7 genome to determine allele frequencies of point mutations across the DHODH gene (presented in Fig. S4).

IV. Junction Identification and Orientation of the DHODH Amplicon

Edges of the DHODH amplicon were estimated from WGS read coverage histograms (Fig. 3A). To identify the junctions between amplicons as well as neighboring sequences, we used the paired-end information (matching pairs) for reads at the very edges of the amplicons (Fig. 3B and C). Reads that aligned (using Bowtie [68]) at the red/yellow arrows (+/−200 bp) were isolated and their matching pairs (opposite end of the read in reverse direction, blue/green arrows (Fig. S6A)) were aligned across the entire 3D7 reference genome. The percentage of total reads that align to either side of the amplicon junction were tallied and mapped for each clone (Fig. S6B).

Accession Numbers

Plasmodium falciparum Dihydroorotate dehydrogenase (DHODH), Genbank Accession Number: AB070244.

Figure 1: **Schematic history of clones selected for varying DSM1 resistance.**

Color codes are conserved in all figures. Clones used in the drug removal experiments are shown with a “*” and underlined clone names were Illumina sequenced. The round 2 naming convention is as follows: first position, letter of the round 1 clone from which it was derived; second position, number of parasites used in selection (5 refers to 105, 7 refers to 107); third position, concentration of DSM1 used during selection (μM, structure shown as inset on right); last position, refers to clone number.

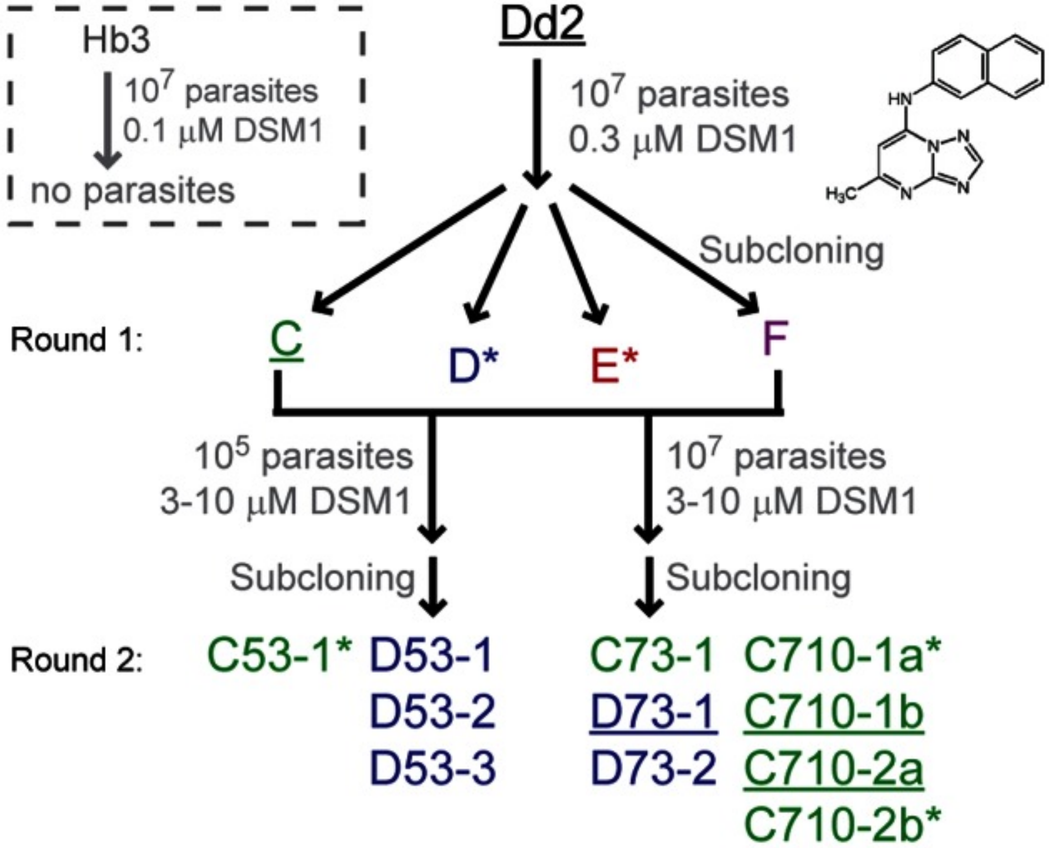


Figure 2: Genes within DHODH amplicons from round 1 clones.

A. Mid-density microarray results of round 1 clone C (other clones, Fig. S1) showing a 70 kb amplicon at the beginning of chromosome 6 (DHODH amplicon). B. Gene coverage (spotted microarray) of the DHODH amplicon (D (blue); C (green); E (red); F (magenta). Average log2 ratios are calculated from 3 experimental replicates over 1–3 probes per gene. There were no other significant amplifications detected anywhere in the genome (significance cut off >1.5 fold change, FDR <10%). Gene numbers 1–25 correspond to those listed in Table S7. C. Summary of DHODH amplicon size (both spotted and mid-density microarrays). The DHODH target gene (no. 19) is depicted in red, A/T tracks at amplicon junctions are indicated by a grey circle (black outline, shared junction; red outline, junction within introns). The amplicon boundaries of each clone were verified using qPCR (Fig. S8). D. Confirmation of DHODH copy number by qPCR. Front and rear primers (Table S12) were used to detect the DHODH gene. Values are relative to Dd2 (grey) and normalized against seryl t-RNA synthetase (PF07_0073) copy number. Error bars depict standard error. Significance was determined against Dd2 (*, p value<0.05 and **, p value<0.005).

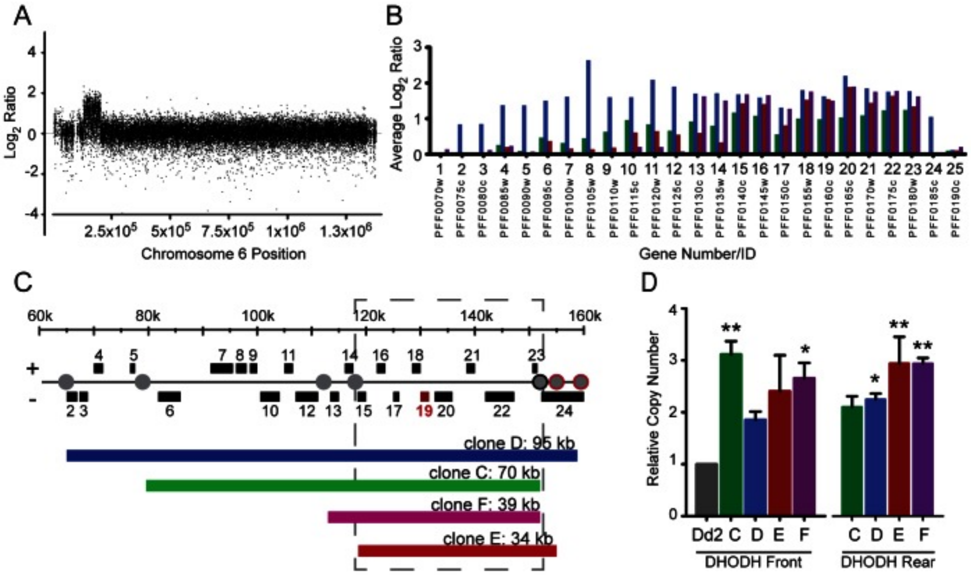


Figure 3: Whole genome sequencing to characterize junctions of DHODH amplicon units. A. Histograms of normalized read coverage comparing single-copy Dd2 (grey), round 1, and 2 clones (C-derived, green; D-derived, blue). The scale depicts chromosome 6 position and boxes below represent ORF locations. All histograms are plotted on the same scale; increases in height correlate to increased number of reads from amplifications. Position of DHODH (vertical grey highlight), junction regions 1 and 2 for C (vertical green highlight) and D (vertical blue highlight)-derived clones. B and C. Mapping of DHODH amplicon junctions using whole genome sequencing data from C (panel B) and D (panel C)-derived clones. Reads that aligned on either side of the junctions were queried for their paired-end alignments to determine amplicon orientation (Fig. S6). Red arrows; reads that align upstream of the amplicon edge, Yellow arrows; downstream of amplicon edge, Green/blue arrows; within the amplicon. Junction positions for each clone are indicated below histograms. *, reads from this position also map to position ~1300500 (chromosome 6). D. Schematic of the tandem head-to-tail orientation of the various clones and the number of amplified regions (arrows) in the Dd2 parent and each resistant clone.

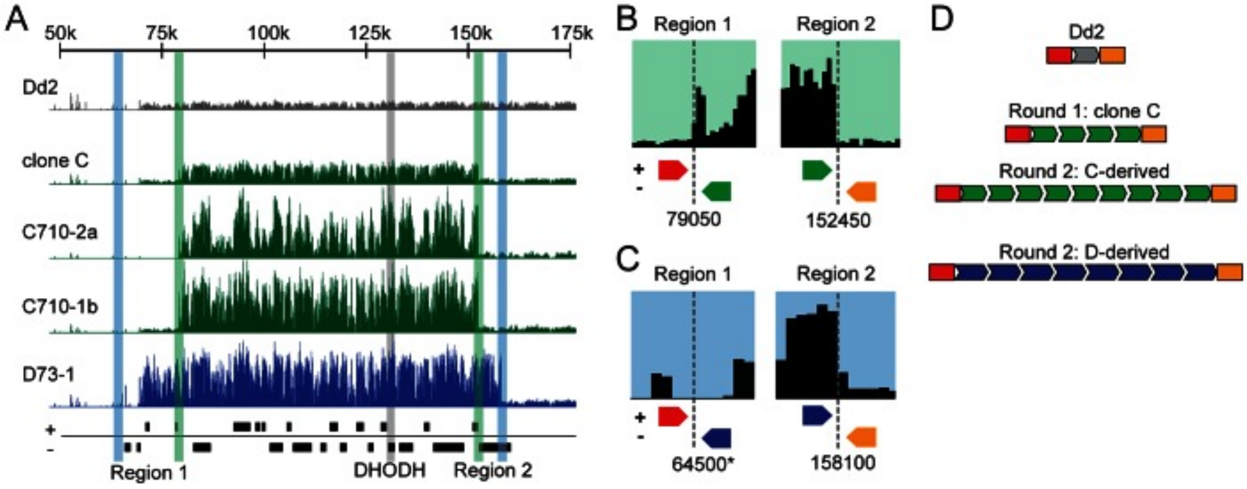


Figure 4: Parallel increases and decreases in DSM1 sensitivities and DHODH amplicons. A. Changes in DSM1 sensitivity of Dd2 (open circle) and C-derived clones (C53-1 (square) and C710-1b (triangle)) from a representative dose response experiment (EC50 values and full list of experiments, Table S9). B. DHODH qPCR showing further amplification in round 2 clones (round 1 C clone included for comparison and used to determine significance). Values are relative to Dd2 (Table S6) and normalized against PF07_0073 and MAL13P1.435. C. Mid-density microarray result from a representative round 2 clone C53-1 (relative to Dd2) showing an increased log2 ratio of the DHODH amplicon on chromosome 6 (mean log2 ratios for all comparisons, Table S6). D and E. DSM1dose response curve of C53-1 (D) and C710-1a (E) populations after growth in the presence (solid line, filled shape) or in the absence (dotted lines, open shapes) of DSM1 for 0 (circle), 1 (square), 2 (triangle), and 3 (diamond) months. F and G. DHODH qPCR analysis of C53-1 (F) and C710-1a (G) populations after 0 to 3 months without DSM1 (black, significance was determined against the +DSM1 population) and −3 month clones (grey, DSM1 removal (DR) clones 1–4). Values are relative to Dd2 and DR clone 1 was used to determine significance. All panels: error bars depict standard error and **, p value<0.005; ***, p value<0.0005. H and I. Tuning of the DHODH amplicon: parental amplicon C clone (left, H and I); round 2 amplicons C53-1 (middle, H) and C710-1a (middle, I); DSM1 removal amplicons C53-1 DR clone 4 (right, H) and C710-1a DR clone 4 (right, I). All were compared to Dd2, except DR clones (right, H and I) were compared to the parental C clone.

Figure 4 cont.

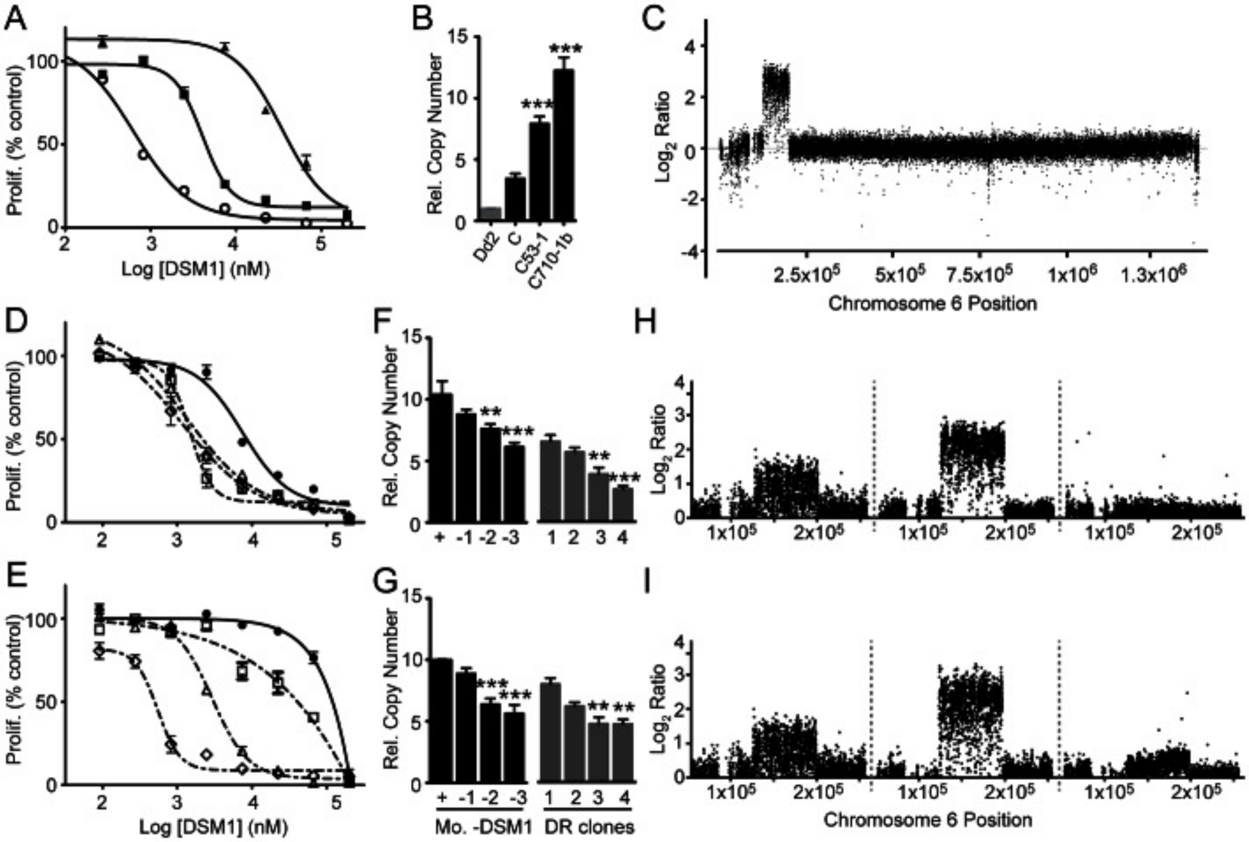


Figure 5: Model of the two-step process that P. falciparum uses to acquire DNA amplifications.

In Step 1, random A/T tracks (grey circles) throughout the haploid genome (black line) initiate a short-homology mediated pathway through presumably either the generation of DNA double-strand breaks due to polymerase pausing or enzymatic action on DNA that is free of histone interactions (see Discussion). In our independent selections, the randomness of the duplication of the genome surrounding DHODH (light blue rectangle) is emphasized by the positions of various initiating A/T tracks (vertical dotted lines) and the generation of differently sized founder amplicons (red, purple, green, and blue bars). The amplicon junction (red line) appears to be generated from uneven “stitching” of the initiating A/T tracks from either side of the amplicon and not simply addition. In Step 2, larger stretches of homology (example green bar) likely trigger homologous recombination-like pathways in the parasite which act to conserve the original beneficial amplicons from Step 1. Long term, the condition of pseudo-plyploidy could allow the generation of mutations (yellow star) across the amplicon, which partial and complete de-amplification could resolve over time.

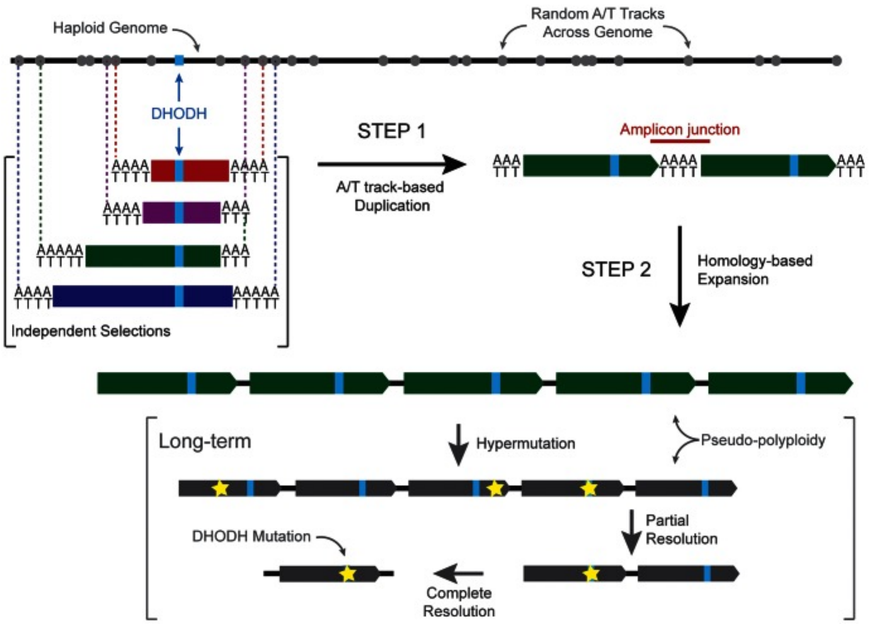


Figure S1: CGH results (mid-density microarrays) for chromosome 6 of round 1 clones. The log2 ratio plot for the C clone is displayed in Fig. 2A.

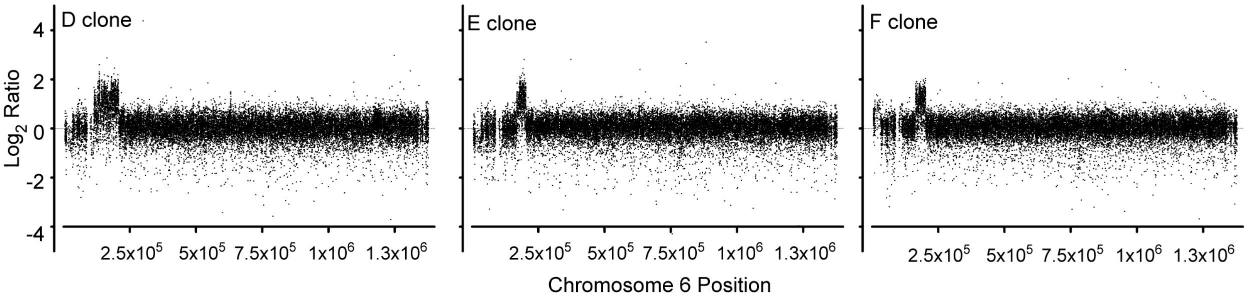


Figure S2: A. DHODH mRNA levels for each round 1 clone (C, green; D, red; E, blue; F, magenta) as determined by expression microarrays. Log2 ratios from DHODH probes (on spotted DNA microarrays) were converted to relative expression levels and mean values (from 2 separate probes hybridized in triplicate) are plotted with error bars (SEM). One-way ANOVA analysis confirms that the difference between clones is not significant. B. DHODH protein levels for each round 1 clone as determined by Western blot analysis. Although only a small region of the blot is shown, no other bands besides that for DHODH (~65 kD) were visible. A portion of the coomassie stained gel from the same experiment is included as the loading control.

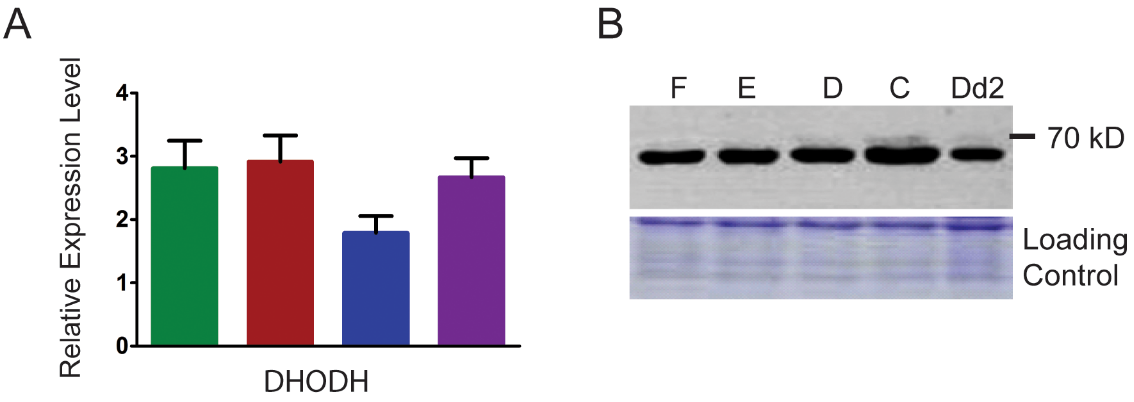


Figure S3: Targeted DHODH sequencing of round 1 and 2 clones. A consensus sequence was generated following assembly of 7 contigs across the 1.7 kb gene for each clone (sequencing primers listed in Table S12) and then compared via ClustalW alignment (Geneious Pro 5.5.6). The green bar displays 100% identity between sequences.

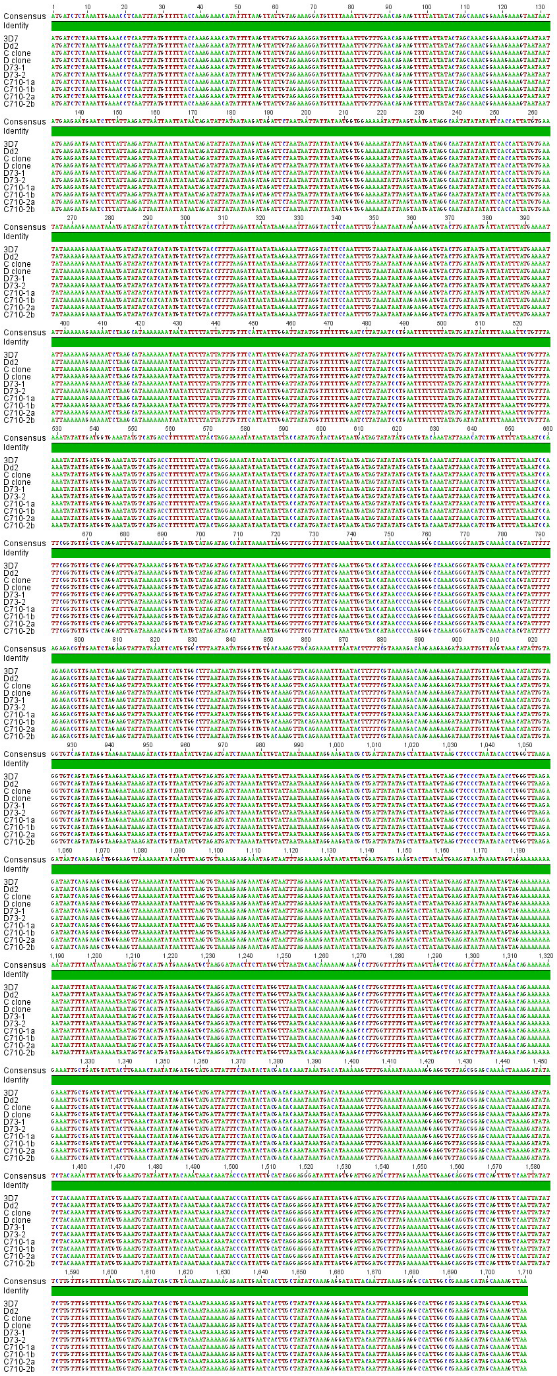


Figure S4: The identification of point mutations across amplified DHODH. Whole genome sequencing reads that aligned to the DHODH gene are scanned for mismatches against the reference 3D7 genome using the Integrated Genome Browser. Histogram bars are colored (green/red) if the allele frequency of a mutant base is >0.05 (1 in 20 reads), otherwise histogram bars are colored in grey. Y axis is presented in a log scale (axis height for each clone is depicted to the right of the plots). Due to the deep coverage of this region of the genome (>50-fold at all nucleotide positions, Table S3), we can confidently conclude that there are no hidden mutations within the amplified DHODH gene. Colored bars in intergenic regions just upstream and downstream of this gene were judged to be sequencing errors based on neighboring repetitive bases.

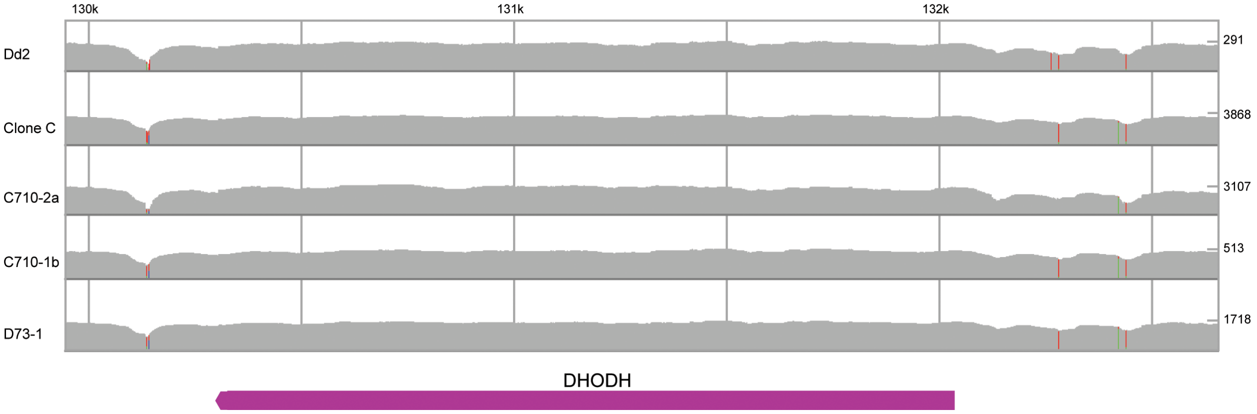
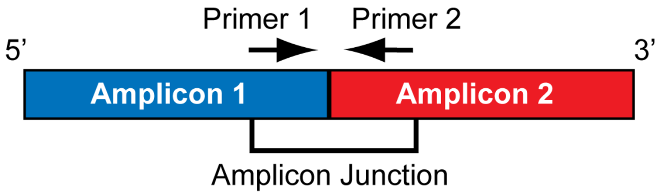


Figure S5: Summary of results for PCR/sequencing of round 1 amplicon junctions. A. Schematic of approach to PCR across the junction of two amplicons in the same orientation. Primers 1 and 2 vary depending on the clone (primer sequences are listed in Table S12). B. Summary amplicon junctions from each round 1 clone. Presence of the junction is unique to clones with the chromosome 6 amplicon. A1; sequence from the 3′ end of amplicon 1, A2; sequence from the 5′ end on amplicon 2. Sequences were compiled and found to be identical between selected colonies of each round 1 clone (see Text S1) and therefore, only 1 sequence per clone is represented. The starting position (*), A1, and A2 sequence is based on 3D7 genome from PlasmoDB (http://plasmodb.org/plasmo/) (although this data may not exactly match the Dd2 genome, preliminary investigation of Dd2 sequence from the Broad Institute (http://www.broadinstitute.org/annotation/genome/plasmodium_falciparum_spp/MultiHome.html) indicates that this data is reasonably accurate). In all cases (except for the A/T track from clone E (**) which contains 2 T's), "A" followed by a number represents an uninterrupted track adenines of the specified length. For example, "A27" indicates the position of a track of adenines 27 bp long.

Figure S5 cont.

A



B

Clone	Description	Starting position*	Sequence
C	Junction	-	CGGATGCTCATCACAAAAG(A27)TAATATATAATAAAATAATC
	A1	152432	CGGATGCTCATCACAAAAG(A36)……………………………….
	A2	79066	……………………………….. (A29)TAATATATAATAAAATAATC
D	Junction	-	AAGCACCTTTCCCCCCAAC(A21)CTTAAGAAATTAACATATA
	A1	158079	AAGCACCTTTCCCCCCAAC(A28)…………………………………
	A2	64524	…………………………………(A38)CTTAAGAAATTAACATATA
E	Junction	-	TGAAATCTGGAAAGACGAG(A19)TTAAATACATAATGGATAT
	A1	152876	TGAAATCTGGAAAGACGAG(A26)**…………………………….
	A2	118425	…………………………………(A21)TTAAATACATAATGGATAT
F	Junction	-	CGGATGCTCATCACAAAAG(A27)GCAACAAAAAAAAAATGTT
	A1	152432	CGGATGCTCATCACAAAAG(A37)……………………………….
	A2	113523	…………………………………(A32)GCAACAAAAAAAAAATGTTTT

Figure S6: WGS-mediated DHODH amplicon junction investigation. A. Mapping of DHODH amplicon junctions from WGS paired-end reads. Reads that aligned +/−200 bp surrounding the junctions were queried for their paired-end alignments to the 3D7 reference genome. Panel depicting C clone junction shown for reference (D clone junction, Fig. 3C). B. Quantitation of the matching pairs from the initial reads mapping to the windows (A, B, C, D) diagramed in panel A. For Dd2, the matching pair always aligns to the neighboring sequence (A maps to B, B maps to A and C maps to D, D maps to C). For clones containing the chromosome 6 amplicon, the matching pair predominantly aligns to the opposite end of the amplicon (i.e. the paired-end reads of region B align to region C and vice versa) indicating a tandem head-to-tail arrangement. Unaligned reads (white box) represent those likely to span the amplicon junction; properties such as low complexity and strain differences limited their alignment to the 3D7 reference genome. *, no initial reads map to this loci due to low genome sequencing coverage. **, alignment is not unique, reads map to another position on chromosome 6 (~1,300,000). ***, no initial reads map to this loci because of mappability (not unique sequences).

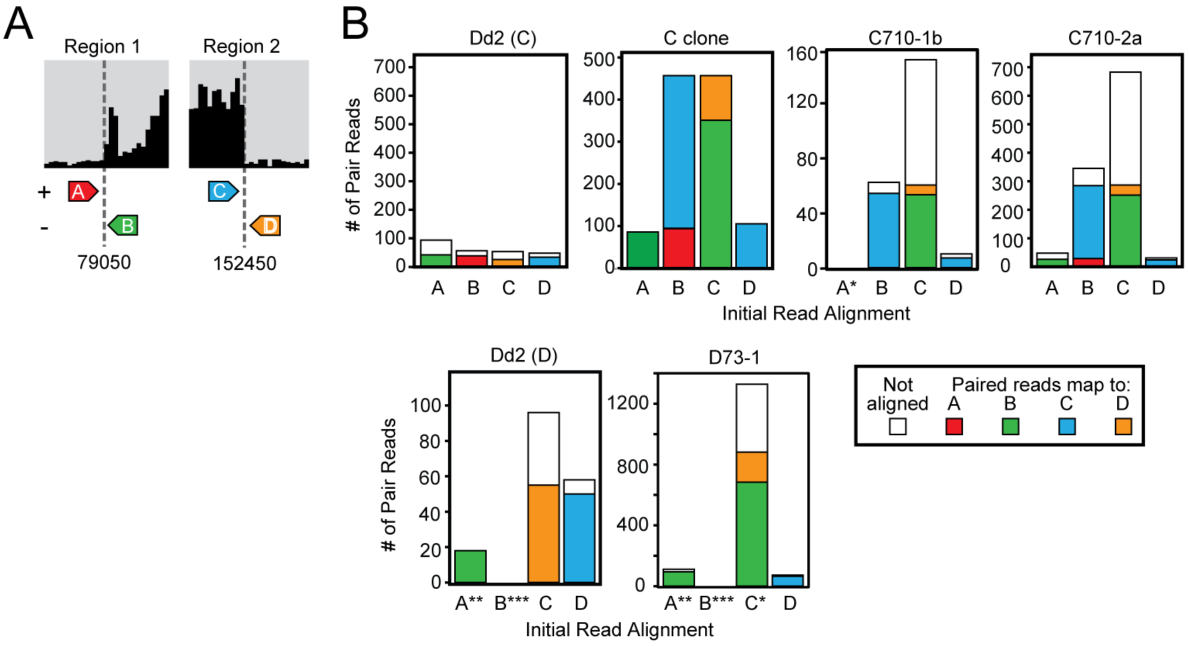


Figure S7:

In vitro growth assessment of round 1 and 2 clones in the presence and absence of DSM1. A. Growth of DSM1 resistant clones was compared to Dd2 (open circle, solid line) over 6 days in multiple independent experiments. Values from these experiments were combined to determine an overall trend for each set of clones: round 1 (clones C and D) solid square, dashed line; round 2–3 μM resistance (C53-1, D53-3, D73-1) solid diamond, dashed line; round 2–10 μM resistance (C710-1a, 1b, 2a, 2b) solid triangle, dashed line. Percent parasitemia values were normalized to the maximum growth of the Dd2 clone in each experiment and plotted as Mean Normalized Parasitemia. Error bars indicate SEM. Beginning on day 4, there is a statistically significant difference in the parasitemia of round 2 clones compared to Dd2 indicating a growth defect (two-way ANOVA, day by clone interaction F (15,80)=9.162 and p<0.001, followed by Bonferroni posttests). Round 2–3 uM and –10 uM clones on average grow 54±8 and 50±16% slower compared to wild-type Dd2 clones, respectively. B. Growth of C53-1 (left plot, triangle) and C71-1a (right plot, diamond) during DSM1 removal experiments. Parasites were cultured in the presence (closed shape, solid line) or absence (open shape, dotted line) of DSM1 for 45 days before growth was measured as in (A) for an additional 14 days and plotted as Mean Normalized Parasitemia (normalization was performed against the maximum growth of the respective – DSM1 clone). Significance could not be determined because only a single value was measured for each time point. While there is no difference in growth between C53-1 ± DSM1, C710-1a (and 2b, data not shown) regains 56% of its growth rate. Growth of Dd2 (grey line) was included for comparison.

Figure S7 cont.

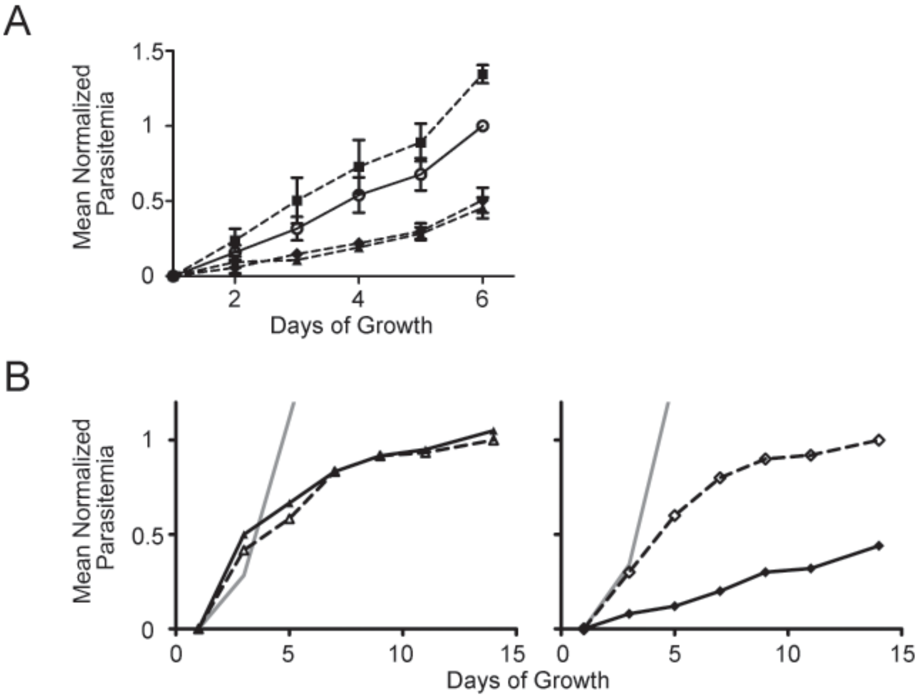


Figure S8: qPCR analysis of copy number of various genes across the amplified region of chromosome 6 (italicized genes in Table S7, primers in Table S12). C, green; D, red; E, blue; F, magenta. All values are relative to Dd2 (grey), normalized against seryl t-RNA synthetase copy number (data normalized to the 18 s ribosomal RNA gene displayed similar results), and determined from multiple experiments. Error bars depict standard error. Significance was determined against Dd2 (***, p value<0.0005).

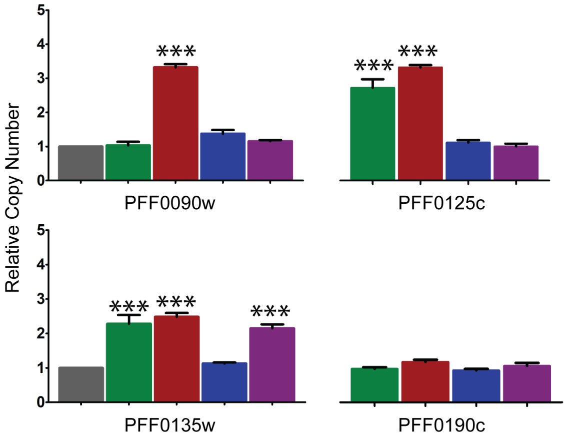


Figure S9: Characteristics of newly selected (NS) DSM1 resistant clones. A. EC50 plots comparing Dd2 (open circle, EC50 value: 0.1±0.01 μM) to uncloned parasites selected with 0.3 μM DSM1 (closed circle, EC50 value: 0.3±0.03 μM). Parasite proliferation was measured in triplicate using the hypoxanthine uptake assay and expressed as a percentage of total radioactivity count from the DMSO control. Error bars depict standard error. B. qPCR analysis of DHODH copy number in two NS clones (clone 1 mean 5.6±0.4, clone 2 mean 4.8±0.4). Significance was determined relative to Dd2 (***, p value<0.0005).

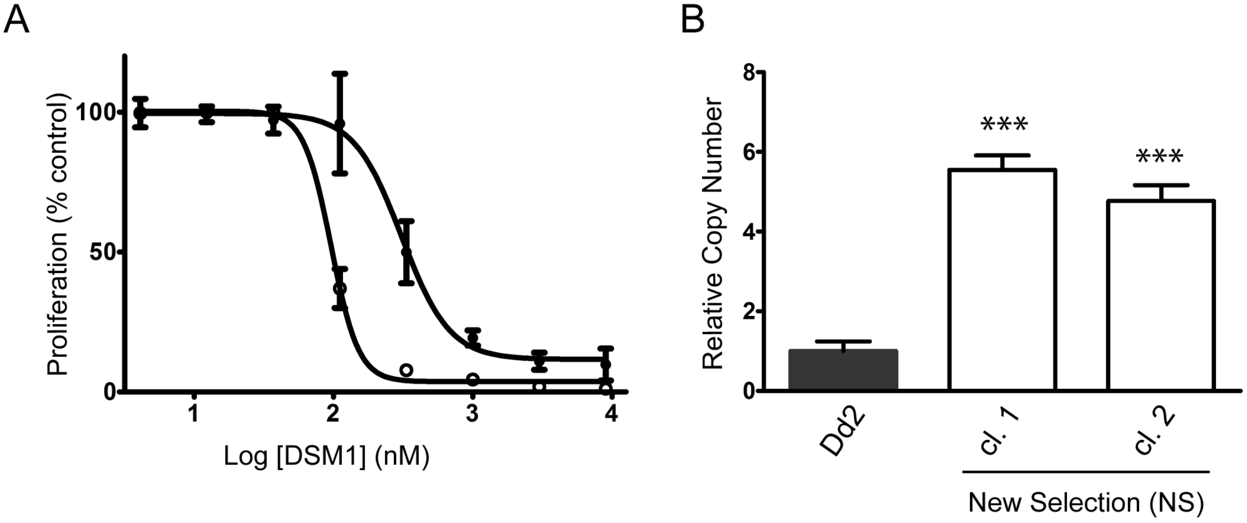


Table S1: Summary of initial DSM1 challenge results. Dd2 (DSM1 EC50 of 0.2 *μ*M) and HB3 (DSM1 EC50 of 0.06 *μ*M) infected erythrocytes were challenged with various concentrations of DSM1 to determine the highest level of achievable resistance. Population sizes of 102–106 Dd2 or Hb3 parasites were also tested but not able to develop resistance to 0.1 *μ*M or higher concentrations of DSM1 (unpublished data).

		Flasks Positive/Flasks Setup (Days to observe parasites ±SD)	
[DSM1] (μM)	**Initial Population**	**Dd2**	**Hb3**
No Drug	10^1	3/3 (13±2)	3/3 (28±0)
0.1	10^1	3/3 (39±0)	0/3
0.1	10^7	3/3 (39±0)	0/3
0.3	10^1	0/3	0/3
0.3	10^7	3/3 (48±7)	0/3
1.0	10^1	0/3	0/3
1.0	10^7	0/3	0/3
3.3	10^1	0/3	0/3
3.3	10^7	0/3	0/3
10	10^1	0/3	0/3
10	10^7	0/3	0/3

Table S2: Round 1 selections. 107 Dd2 parasites were plated over 24 wells (4 replicates) and challenged with 0.3 μM DSM1 (Note: populations of <107 parasites were not able to survive treatment with this DSM1 concentration). In total, 8 wells were positive for resistant parasites (round 1 clones). Four of these wells were randomly selected for DSM1 EC50 determination and sub-cloning. One sub-clone of each round 1 clone was selected for further analysis. Nd, EC50 not determined.

Parent clone/ Replicate Number	Round 1 Clone*	Sub-clone	Days to Detection	EC_{50} ± 95% CI (μM)
Dd2^{sensitive}	--		13	0.2±0.0
1	B3**		93	--
2	C12		85	1.0±0.1
		C12sC8 (C)	-	1.1±0.1
	D9		36	0.9±0.2
		D9sD5 (D)	-	0.9±0.2
	E1		52	Nd
	E4		66	Nd
3	E10		17	1.2±0.1
		E10sD6 (E)	-	0.9±0.2
	F2		22	Nd
	F4		88	1.0±0.1
		F4sH12 (F)	-	0.9±0.1
4	none		--	--

*The clone name is based on the coordinates of the well in which it was isolated from the selection plate.
**This clone did not grow in 10 ml flask in presence of 0.3 μM DSM1.

Table S3: Whole genome sequencing coverage rates of various regions of interest across the P. falciparum genome. Overall very deep coverage was achieved in all clones except C710-1b clone (designated "*"). Whole genome rates are considerably lower than those for the DHODH gene presumably due to the inclusion of intergenic regions in this data set where base composition may limit the unique alignment of many reads. Coverage rates within clone C and D amplicon boundaries are included to emphasize the very deep coverage of these areas and thus, our confidence in the lack of mutations across these regions. Nd, not determined.

		Covered by Number of Reads:					
Region of Genome	Clone	>1	>5	>10	>20	>50	>200
Whole Genome	Dd2	88.3%	86.2%	84.7%	81.6%	77.1%	Nd
	C	88.6%	87.5%	86.9%	85.9%	83.2%	Nd
	D73-1	88.0%	86.4%	85.3%	83.5%	77.6%	Nd
	C710-1b*	83.9%	66.9%	53.8%	32.6%	1.3%	Nd
	C710-2a	87.7%	84.9%	82.4%	77.2%	61.6%	Nd
DHODH Gene	Dd2	100.0%	100.0%	100.0%	100.0%	100.0%	30.2%
	C	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	D73-1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	C710-1b*	100.0%	100.0%	100.0%	100.0%	100.0%	81.0%
	C710-2a	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Clone C Boundaries	Dd2	98.9%	98.0%	96.8%	94.2%	83.7%	7.5%
	C	99.4%	99.1%	98.9%	98.8%	98.5%	95.5%
	C710-1b*	98.9%	97.6%	95.8%	92.2%	79.2%	33.5%
	C710-2a	99.4%	98.9%	98.8%	98.6%	97.4%	89.1%
Clone D Boundaries	Dd2	95.1%	93.7%	91.9%	89.1%	78.3%	6.6%
	D73-1	95.9%	95.3%	95.1%	94.7%	94.0%	90.3%

Table S4: Summary of mutations in the DHODH amplicon. In order to find low frequency mutations in an amplified region, positional nucleotide frequencies were identified by comparing Illumina reads from resistant clones and Dd2 that cover the amplicons on chromosome 6. This result of this analysis across the DHODH gene is also summarized in Fig. S4. DHODH (PFF0160c) is the target of DSM1. (−) no SNPs detected.

	Round 1	Round 2		
Exon	C	C710-1b	C710-2a	D73-1
PFF0095c	-	-	-	SNP85172 (T->A) SNP85177 (A->T) SNP85188 (T->C)
PFF0100w	-	-	-	-
PFF0105w	-	-	-	-
PFF0110w	-	-	-	-
PFF0115c	-	-	-	-
PFF0120w	-	-	-	-
PFF0125c	-	-	-	-
PFF0130c	-	-	-	-
PFF0135w	-	-	-	-
PFF0140c	-	-	-	-
PFF0145w	-	-	-	-
PFF0150c	-	-	-	-
PFF0155w	-	-	-	-
PFF0160c	-	-	-	-
PFF0165c	-	-	-	-
PFF0170w	-	-	-	-
PFF0175c	-	-	-	-
PFF0180w	-	-	-	-

Table S5: SNPs present in >90% of Illumina sequencing reads. The top 5 listed SNPs were validated by PCR and sequencing (Table 1) and shown to preexist in Dd2 (see Materials and Methods section of main paper).

Chr.	Position	AA Change	ID*	Description **	Round 1	Round 2		
					C clone	D73-1	C710-1a	C710-2b
5	214244	Cys->Ser	PFE0245c	Membrane protein	+	+	+	+
6	645035	Asn->Asp	PFF0750w	Cdc-related protein kinase	+	+	+	+
8	738807	Tyr->Asn	MAL8P1_82	Vacuolar sorting protein	+	+	+	+
5	214184	Cys->Ser	PFE0245c	Membrane protein	+	+	(+)	+
14	721985	Lys->Stop	PF14_0173	cAMP binding protein	+	+	-	+
12	151552	Lys->Ile	PFL0130c	Conserved/ unknown function	-	-	+	-
8	902680	Lys->Asn	PF08_0048	Snf2-related CBP activator	-	-	+	-
12	865888	Met->Arg	PFL1045w	Conserved/ unknown function	-	-	-	+
14	500065	Ser->Asn	PF14_0123	Conserved/ unknown function	-	-	-	+
6	404586	Ser->Phe	PFF0470w	Conserved/ unknown function	-	-	-	+

*PlasmoDB gene ID

**Basic gene description based on PlasmoDB functional assignments.

(+) does not qualify for all conditions but SNP is present in C710-1a (covered by only 2 reads (cutoff for filtering is 5 reads))

Table S6: Summary of chromosome 6 amplicon boundaries and DHODH copy numbers for key round 1 and 2 clones. Microarray probe positions were judged from mid-density microarray analysis as the first (Start) and last (Stop) probe that exhibited a log2 ratio >0.3 in the amplicon border region. Mean log2 ratios were calculated across the entire amplified region. Since exact DHODH copy numbers could not be estimated from amplicons exhibiting log2 ratio >0.8, qPCR was employed (mean of values for front DHODH primer set (Table S12) from multiple experiments). Clones from two DSM1 removal (DR) experiments in which CGH analysis was performed are listed underneath the clone in which they were derived. All CGH experiments are pair-wise comparisons against Dd2 genomic DNA (except DR clones are compared to the round 1 C clone). Nd, not determined.

Round	Clone	Microarray Probe Position		True Genome Region*	Mean Log_2 Ratio	DHODH Copy Number	
		Start	Stop			CGH[$]	qPCR (±SE)
1	C	129409[%]	202401	79409-152456	0.8	3	4.0 ± 0.4
2	C53-1	129209	202401	79209-152456	2.4	>4	8.3 ± 0.5
	DR clone 4	*129409*[#]	*202401*	*79409-152456*	-	-	2.7 ± 0.3
	C73-1	129209	202401	79209-152456	2.5	>4	Nd
	C710-1a	129209	202401	79209-152456	2.5	>4	10.0 ± 1.7
	DR clone 3	129209	202401	79209-152456	-	-	4.8 ± 0.4
	C710-1b	Nd	Nd	Nd	Nd	Nd	12.2 ± 0.7
	C710-2b	129209	202401	79209-152456	2.3	>4	11.5 ± 0.5
1	D	114617	208017	64619-158072	0.9	4	3.5 ± 0.3
2	D53-1	114617	208145[&]	64619-158204	2.0	>4	Nd
	D53-2	Nd	Nd	Nd	Nd	Nd	8.6 ± 0.5
	D73-1	114617	208017	64619-158074	1.5	>4	12.3 ± 1.2
	D73-2	114617	208017	64619-158074	2.2	>4	Nd

*The genome region is different from probe position due to a misalignment of microarray probes by ~50kb in this region. This is likely due to the status of the alignment of the *P. falciparum* genome project at the time of microarray design.

[$] Based on NimbleGen log_2 ratio scale: 0.25 to 0.5= 1 additional unit, 0.5 to 0.8= 2 additional units, >0.8= 3+ additional units.

[%] Probe 129409 is 2 probes away from 129209 on the mid-density microarray and does not likely represent a difference in the location of the amplicon junction.

[#] An approximation based on CGH comparison to round 1 clone C (in italics). Mean log_2 ratio was not included because comparisons were made against the round 1 C clone instead of Dd2.

[&] Probe 208145 is 1 probe away from 208017 and does not likely represent a difference in the location of the junction.

Table S7: Description of genes contained within the largest DHODH amplicon. The DHODH target of DSM1 is bolded. Other genes selected for qPCR analysis (see Table S12) are italicized. The smallest amplicon (from clone E) encompasses gene numbers 15 to 23.

Gene Number*	ID**	Description***
1[#]	PFF0070w	PfEMP1 pseudogene
2	PFF0075c	PHISTb exported protein
3	PFF0080c	TRAP-like protein
4	PFF0085w	PHISTa exported protein
5	*PFF0090w*	*Conserved/unknown function*
6	PFF0095c	Conserved/unknown function
7	PFF0100w	ATP-dependent RNA helicase
8	PFF0105w	MYND finger protein
9	PFF0110w	Liver merozoite formation protein
10	PFF0115c	Elongation factor G
11	PFF0120w	Geranylgeranyl transferase
12	*PFF0125c*	*Conserved/unknown function*
13	PFF0130c	Conserved/unknown function
14	*PFF0135w*	*JmjC domain containing protein*
15	PFF0140c	Conserved/unknown function
16	PFF0145w	Conserved/unknown function
17	PFF0150c	Conserved/unknown function
18	PFF0155w	Mitochondrial chaperone BCS-1
19	***PFF0160c***	***DHODH***
20	PFF0165c	Conserved/unknown function
21	PFF0170w	Cation/H+ antiporter (PfCHA)
22	PFF0175c	Conserved/unknown function
23	PFF0180w	Phenylalanyl-tRNA synthetase subunit
24	PFF0185c	Conserved/unknown function
25[#]	*PFF0190c*	*Conserved/unknown function*

*From Fig. 2B and C.

**PlasmoDB gene ID

***Basic gene description based on PlasmoDB functional assignments.

[#]Not included in longest DHODH amplicon (D).

Table S8: Round 2 selections. Populations of parasites were challenged with 1–10 μM DSM1 and scored for positive growth over 48 (for an initial population of 101) or 96 days (all other conditions). Resistant parasites from parental clone C and D were sub-cloned for further analysis (*) but those from clones E and F were not followed further. Nd, not determined.

Parent Clone/ Round 1 Clone	Initial Population	Wells Positive/Wells Setup (Days to observed parasites)		
		1 μM	3.3 μM	10 μM
Dd2$^{\text{sensitive}}$	10^1	0/48	0/48	Nd
	10^5	Nd	Nd	Nd
	10^7	0/96	0/96	Nd
C	10^1	0/48	0/48	0/48
	10^5	Nd	16/72 (24)*	Nd
	10^7	72/72 (9)	72/72 (19)*	7/72 (25)*
D	10^1	0/48	0/48	0/48
	10^5	Nd	1/96*	Nd
	10^7	96/96 (7)	39/96 (18)*	0/96
E	10^1	0/48	0/48	0/48
	10^5	Nd	26/96 (23)	Nd
	10^7	96/96 (7)	96/96 (17)	10/96 (26)
F	10^1	0/48	0/48	0/48
	10^5	Nd	24/96 (17)	Nd
	10^7	96/96 (9)	78/96 (22)	25/96 (25)

Table S9: Summary of EC50 values for round 2 clones. Values were determined using a higher range of DSM1 (0.09–200 μM) than was used for round 1 clones. These concentrations approach saturation for this compound, which may explain the increase in Dd2 EC50 (compared to value in Table S2) and high variability observed in these experiments. Fold increase was calculated against Dd2 from this table (tested at the high range of DSM1). Results from drug removal (DR) experiments are listed underneath the clone in which they were derived from (−1: 1 month without DSM1). Nd, could not be determined.

Clone	Exp. No.	EC_{50} (μM)	95% CI	Mean EC_{50} (μM)	~Fold Increase
Dd2	1	0.3	±0.2	0.5	**-**
	2	0.6	±0.2		
	3	0.3	±0.01		
	4	0.6	±0.02		
C53-1	1	9.5	±2.7	7.2	**15**
	2	4.9	±1.2		
	DR-1	1.5	±0.3	-	3
	DR-2	1.3	±0.8		2
	DR-3	1.7	±0.5		3
C710-1a	1	29	±12	62	**130**
	2	66	±22		
	3	95	±11		
	4	59	±13		
	DR-1	34	±12	-	60
	DR-2	3.2	±4.2		5
	DR-3	1.0	±0.3		2
C710-1b	1	48	±21	85	**180**
	2	122	±93		
C710-2a	1	65	Nd	56	**120**
	2	46	±12		
C710-2b	1	59	±20	53	**115**
	2	52	±10		
	3	45	±14		
	4	59	±20		
	DR-1	25	±10	-	40
	DR-2	1.9	±1.2		3
	DR-3	1.3	±2.3		2
D53-1	1	9.1	±1.9	36	**75**
	2	62	±11		
D53-2	1	65	Nd	65	**140**
D73-1	1	49	±6.8	49	**100**

Table S10: EC50 values for DSM1 resistant round 1 clones D and E after 8 months of continuous culture with or without (*) 0.3 μM DSM1 pressure. Fold-resistance values are included in parentheses for ease of comparison.

	DSM1 EC_{50} (μM) ± 95%CI (Fold-resistance)									
Round 1 Sub-clone	**0 months**		2 months		4 months		6 months		8 months	
D*	**-**		0.5±0.1	(2.5)	0.5±0.1	(2.5)	0.6±0.1	(3)	0.5±0.1	(2.5)
D	**0.9±0.2**	**(4.5)**	0.6±0.0	(3)	1.1±0.1	(5.5)	1.0±0.1	(5)	0.9±0.1	(4.5)
E*	**-**		0.6±0.1	(3)	0.7±0.2	(3.5)	0.5±0.1	(2.5)	0.5±0.1	(2.5)
E	**0.9±0.2**	**(4.5)**	1.0±0.1	(5)	1.5±0.3	(7.5)	0.9±0.1	(4.5)	1.3±0.3	(6.5)

Table S11: Additional antimalarial EC50 determination (±95% CI) for DSM1 resistant clones. Proguanil and 1843U89 target the P. falciparum dihydrofolate reductase enzyme and thymidylate synthase, respectively [72]. 5-Fluoroorotate inhibits pyrimidine biosynthesis in P. falciparum [73], [74]. Artemisinin is currently used clinically and its target remains unidentified (reviewed in [75]). Assay type A is a flow cytometry-based method that involves measurement of parasitemia using SYBR green and assay type B depends on the uptake of radiolabeled hypoxanthine (see Materials and Methods and Text S1). (−), experiment not performed. Nd, could not be determined.

Parasite Clone/ Antimalarial	Proguanil EC$_{50}$ (µM)	5-Fluoroorotate EC$_{50}$ (nM)		1843U89 EC$_{50}$ (nM)	Artemisinin EC$_{50}$ (nM)	
Assay Type	A	A	B	B	A	B
Dd2	12.4 (±0.9)	4.5 (±0.2)	7.9 (±0.3)	290 (Nd)	12.5 (Nd)	11.5 (±1.4)
C	14.8 (±3.6)	5 (±0.4)	-	-	8.3 (±0.8)	-
D	14.8 (±3.4)	5.2 (±0.6)	-	-	7.4 (±1)	-
C53-1	10.3 (±4.4)	3.6 (Nd)	-	-	12.5 (Nd)	-
D73-1	-	-	-	-	4.7 (±0.5)	-
D73-2	7 (±2.1)	-	8.3 (Nd)	244 (±24.1)	-	4.4 (±0.7)
C710-1a	10.3 (±2.7)	3.3 (Nd)	-	-	5 (Nd)	-
C710-2a	17.6 (±3.9)	5.1 (±0.5)	-	-	5.4 (±0.4)	-

Table S12: Summary of primers by experiment

Exp.	Details		Primer Sequence	Product size
	Gene ID	Function		
qPCR	PFF0090w	Unknown	F-CCAAAATGTCAAAACACTATG R-CTGCATTGGCTGAAGCATAAACAG	158bp
	PFF0125c	Unknown	F-CGTCCATGAATGTGAAGAGTGG R-GGATAAGTAGATACAACACTAC	237bp
	PFF0135w	Unknown	F-CAGCCAGGACATACGAAGAGG R-GCATTGCCCTATCTTATCTTG	163bp
	PFF0160c	DHODH Front DHODH Rear	F-TCCATTCGGTGTTGCTGCAGGATTTGAT R-TCTGTAACTTTGTCACAACCCATATTA F-GTGTTAGCGGAGCAAAACTAAAAG R-ATAATTGACAAACTGAAGCACCTG	206bp 158bp
	PFF0190c	Unknown	F-GACGATATTCAGAATGATGTTCAG R- TTTACGATCTTCTTTAACACACC	175bp
	PF07_0073	Seryl t-RNA Synthetase	F-GGAACAATTCTGTATTGCTTTACC R- AAGCTGCGTTGTTTAAAGCTC	142bp
	PFL1155w	GTP cyclohydrolase I	F-AAATATGAGGGGAGTTAAAGAGCA R- TTTAAATTTTCCACAGAAGAGTCA	120bp
	MAL13P1.435	18s Ribosomal RNA	F-ACAATTCATCATATCTTTCAATCGGTA R- GCTGACTACGTCCCTGCCC	69bp
Junction PCR	C junction		*1-CGGATGCTCATCACAAAAGA** 2-TCAAAGGAGAGTCCCAAAGG	~1600bp
	D junction		1-CTGCTGATGGCTAAATTCTCA 2-GACCGTGTGTTGAATAGTTTCTTT	~250bp
	E junction		1-CAGTGAAATCTGGAAAGACGAG 2-TGGATAAACAGGTTGAAAAAGAG	~400bp
	F junction		1-CGGATGCTCATCACAAAAGA** 2-TGTTAATTCCGGGGTTACCTT	~350bp
DHODH Seq.	DHODH-F		CATTTAAGCCCCAAAACATTTTTAC	N.A.
	DHODH-R		GTGATAGATAGCTCCAGTCGATTTC	N.A.
	Seq1		TCATCATATGTATCTGTACCTTTTAAGATT	N.A.
	Seq2		AGCTCCCCTAATACACCTGGGTT	N.A.
	Seq3		TGCAAAACCACGTATTTTTAGAGAC	N.A.
	Seq4		TATATATATATTTTTTTTTTTTTGCGC	N.A.
	Seq5		GCCCTTGGTTTTTGTTAAGTTAGCTCC	N.A.
	Seq6		TCTGTAACTTTGTCACAACCCATATTA	N.A.
SNP Validation	PFE0245c	Position 214184***	F-TCCTCTTCTTTTCTACATGCTACATC R- TAATAAGAATAGGTGGAGACCTTTTTG	~1300bp
	PFE0245c	Position 214244***	F-TCCTCTTCTTTTCTACATGCTACATC R- TAATAAGAATAGGTGGAGACCTTTTTG	~1300bp
	PFF0750w	Position 645035	F-GCATCATCATAAAAGTCATGCAA R- AATGCATGCAGCTGACCATA	~400bp
	MAL8P1.82	Position 738807	F-CGTTCGAAATTAATTCCTTCCA R- GAAAAGCCTCCAAAAGGGATA	~1000bp
	PF14_0173	Position 721985	F-AAGGCCAAAATTTGTCATCTG R- CAGTATCGATTATTGCCACTGC	~700bp

*Primer number refers to the position of the primer on Fig. S5A. All primers numbered 1 are the forward direction and primers numbered 2 are the reverse direction for the PCR reaction.

**The same primer 1 was used because exact 3' junction was predicted by microarray data for C and F clones.

***These SNPs were amplified using the same primers and sequenced in the same reaction since they are only 60 bp apart.

N.A. not applicable

Table S13: Copy number assessment of the chromosome 5 and 12 amplicons in DSM1 resistant clones using microarray and qPCR.

Round	Clone	Chrom. 5 amplicon[*]		Chrom. 12 amplicon[*]		
		Mean log$_2$ ratio[#]	Copy number[$]	Mean log$_2$ ratio	Copy number	qPCR[%]
-	Dd2	-	3[&]	-	2[&]	1.6
1	C	0.2	3	0.5	4	4.4
2	C53-1	0.0	3	0.3	3	1.5[@]
	DR clone 4	*0.0*	*3*	*0.5*	*4*	*Nd*
	C73-1	0.2	3	0.6	4	3.3
	C710-1a	0.1	3	1.2	5	4.2
	DR clone 3	*0.0*	*3*	*-0.7*	*1*	*Nd*
	C710-2b	0.1	3	0.5	3	3.8
1	D	-0.2	3	-0.2	2	1.8
2	D53-1	-0.4	2	0.7	4	4.3
	D73-1	-0.3	2	0.0	2	2.1
	D73-2	-0.4	2	0.0	2	1.8

[*] Boundaries identified from WGS studies: Chromosome 5 (888060-970427) and chromosome 12 (971307-976534)

[#] All clones were compared to Dd2 using CGH (except DR clones (*italics*) were compared to the parental C clone).

[$] Approximate copy number was calculated based on NimbleGen log$_2$ ratio scale: 0.25 to 0.5= 1 additional unit, 0.5 to 0.8= 2 additional units, >0.8= 3+ additional units on top of the number of copies already in Dd2.

[%] Copy number of this region was measured by qPCR using primers against GTP cyclohydrolase (PFL1155w, Table S9). Values are relative to parasite clone FCR3 which has a single copy of this region of chromosome 12 [56].

[&] Levels of the chromosome 5 and 12 amplicons in Dd2 were estimated from WGS studies and qPCR of PFL1155w respectively.

[@] This clone may have lost copies following subsequent rounds of culture.

Supplementary Materials and Methods

Expression Analysis. Parasites were synchronized with 5% sorbitol for two successive cycles and then total RNA was isolated from cultures containing predominantly trophozoite stage using the RNAqueous kit (Ambion). cDNA was synthesized in the presence of aa-dUTP and then coupled to either Cy3 (DSM1-sensitive Dd2) or Cy5 (DSM1-resistant clones) [1,2]. Hybridizations of each pair to spotted DNA microarrays were performed at 62°C for 16-18h, after which, microarrays were washed, dried, scanned, and analyzed as described for CGH.

Determination of DHODH Protein Levels. P. falciparum DHODH was expressed and purified as the N-terminally truncated protein (to remove the membrane spanning domain) as previously described [3,4] and protein (20 mgs) was supplied to Affinity BioReagents (Golden, CO, USA) in 1 ml buffer (100 mM Hepes pH 8.0, 300 mM NaCl, 15% glycerol, 15% triton-reduced) for the generation of rabbit antibodies. Round 1 DSM1 resistant parasite were harvested and run on an SDS-PAGE gel for Western blot analysis. Rabbit anti-recombinant DHODH (1:7000) and goat anti-rabbit horseradish peroxidase (Jackson Laboratories, 1:10,000) were used as primary and secondary antibodies, respectively.

Targeted DHODH Sequencing. Genomic DNA was purified from each DSM1 resistant clone and the sensitive Dd2 clone using the method described in the paper. The DHODH gene was PCR amplified using DHODH-F and DHODH-R (Table S12). The optimized protocol, giving a single amplified product, was found to be 96oC for 3 min, 30 rounds of 96oC for 45 sec, 53oC for 1.5 min, and 68oC for 2 min, with a final extension of 68oC for 10 min. Amplified product was PCR-purified (Qiagen) and dideoxy-sequenced (Eurofins MWG Operon) using 6 additional primers (listed in Table S12) designed to walk across the entire gene with significant overlap.

Consensus sequences were compiled for each clone and compared using Geneious Pro 5.5.6 (Fig. S3).

EC50 Determination by SYBR Green Assay. A parasite solution at 0.5% parasitemia (0.5% hematocrit) from the clone of interest was plated into a 96-well culture plate. An appropriate range of antimalarial concentrations were then added to the parasites (final 0.5% DMSO). Each concentration was performed in triplicate and included solvent-only and uninfected red blood cell controls. After incubating for ~72 hours, the majority of media was removed from each well (leaving red blood cells settled at the bottom). An equal volume of 2X SYBR green (diluted in PBS, Sigma Aldrich) was then added to each well and incubated at room temperature for 20 min before adding cold PBS up to 200μl. 96-well plates were stored at 4°C until parasitemia measurement was performed using the Accurri C6 flow cytometer with a CSampler robotic arm (BD Biosciences). 50,000 counts were collected from each well and gates to exclude debris and aggregates were set similar to [5]. Parasite proliferation in each test well was expressed as a percentage of the solvent control well. EC50 values were fit using the GraphPad PRISM software, according to the equation: Y = Bottom + (Top-Bottom) / (1+10((LogEC50 - X) * HillSlope)).

PCR and Sequencing of the DHODH Amplicon Junction. Genomic DNA was isolated from clones as described in the paper for microarrays using the DNeasy kit (Qiagen). Using unique primers sets for each round 1 clone (Table S12), we amplified the amplicon junction using the following protocol: 94°C 45 sec, 52-55°C 45 sec, 72°C 1 min. PCR products were cloned into the Topo-TA vector (Invitrogen) and transformed into bacteria. Single colonies were isolated (~4 per round 1 clone) and grown for minipreps (Qiagen) prior to di-deoxy sequencing using TOPO-TA-specific M13F/R primers.

Growth Assessment. Due to experimental variability, growth of DSM1 resistant clones were assessed for 2 to 4 parasite clones in 2 to 3 independent experiments and data were combined to determine an overall trend for each group of resistant parasites (i.e. round 1, round 2- 3 μM, and round 2- 10 μM). Percent parasitemia was recorded daily from thin smears, adjusted based on dilution of the culture, and normalized to the maximum growth of the Dd2 clone in each experiment. SEM was determined from the mean of multiple experiments and significance was determined by two-way ANOVA statistical test followed by a Bonferroni posttest. The relative growth rate was calculated by normalizing the average slope of the linear fit of percent parasitemia for each clone to the Dd2 values in each experiment. In the case of DSM1 removal experiments, parasitemia measurements began after 45 days of growth in the absence of DSM1 and normalization was performed against the maximum growth of the respective – DSM1 clone (i.e. growth of C53-1 +DSM1was normalized to growth of C53-1 –DSM1).

Quantitative PCR of other genes contained within the DHODH amplicon. The same qPCR protocol that was used to amplify DHODH was used for PFF0135c and PFF0190c (primer sequences are listed in Table S12). For PFF0090w and PFF0125c, an annealing temperature of 53oC was required. We performed melt curves, standard curves, and analysis as described in the main body of the paper.

Abbreviations

DHODH, dihydroorotate dehydrogenase; gDNA, genomic DNA; CGH, comparative genomic hybridization; WGS, whole genome sequencing; qPCR, quantitative PCR; SNP, single nucleotide polymorphism; NS, newly selected; Nd, not determined; DR, drug removal.

Supplementary References

1. Ganesan K, Jiang L, Rathod PK (2002) Stochastic versus stable transcriptional differences on Plasmodium falciparum DNA microarrays. International Journal for Parasitology 32: 1543-1550.

2. Ganesan K, Ponmee N, Jiang L, Fowble JW, White J, et al. (2008) A Genetically hard-wired metabolic transcriptome in Plasmodium falciparum fails to mount protective responses to lethal antifolates. PLoS Pathog 4: e1000214.

3. Phillips MA, Gujjar R, Malmquist NA, White J, El Mazouni F, et al. (2008) Triazolopyrimidine-based dihydroorotate dehydrogenase inhibitors with potent and selective activity against the malaria parasite Plasmodium falciparum. Journal of Medicinal Chemistry 51: 3649-3653.

4. Gujjar R, Marwaha A, El Mazouni F, White J, White KL, et al. (2009) Identification of a metabolically stable triazolopyrimidine-based dihydroorotate dehydrogenase inhibitor with antimalarial activity in mice. Journal of Medicinal Chemistry 52: 1864-1872.

5. Malleret B, Claser C, Ong AS, Suwanarusk R, Sriprawat K, et al. A rapid and robust tri-color flow cytometry assay for monitoring malaria parasite development. Sci Rep 1: 118.

References

1. White NJ (1992) Antimalarial drug resistance: the pace quickens. J Antimicrob Chemother 30: 571–585

2. Sa JM, Chong JL, Wellems TE (2011) Malaria drug resistance: new observations and developments. Essays Biochem 51: 137–160

3. Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, et al. (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet 379: 413–431

4. Hay SI, Okiro EA, Gething PW, Patil AP, Tatem AJ, et al. (2010) Estimating the global clinical burden of Plasmodium falciparum malaria in 2007. PLoS Med 7: e1000290.

5. WHO (2012) World Malaria Report.

6. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, et al. (2009) Artemisinin resistance in Plasmodium falciparum malaria. N Engl J Med 361: 455–467

7. Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, et al. (2008) Evidence of artemisinin-resistant malaria in western Cambodia. N Engl J Med 359: 2619–2620

8. Thanh NV, Toan TQ, Cowman AF, Casey GJ, Phuc BQ, et al. (2010) Monitoring for Plasmodium falciparum drug resistance to artemisinin and artesunate in Binh Phuoc Province, Vietnam: 1998–2009. Malar J 9: 181.

9. Rogers WO, Sem R, Tero T, Chim P, Lim P, et al. (2009) Failure of artesunate-mefloquine combination therapy for uncomplicated Plasmodium falciparum malaria in southern Cambodia. Malar J 8: 10.

10. Phyo AP, Nkhoma S, Stepniewska K, Ashley EA, Nair S, et al. (2012) Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. Lancet doi:10.1016/S0140-6736(12)60484-X

11. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, et al. (2000) Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. Molecular Cell 6: 861–871

12. Cowman AF, Morry MJ, Biggs BA, Cross GA, Foote SJ (1988) Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 85: 9109–9113

13. Peterson DS, Milhous WK, Wellems TE (1990) Molecular basis of differential resistance to cycloguanil and pyrimethamine in Plasmodium falciparum malaria. Proceedings of the National Academy of Sciences of the United States of America 87: 3018–3022

14. Wilson C, Serrano A, Wasley A, Bogenschutz M, Shankar A, et al. (1989) Amplification of a gene related to mammalian mdr genes in drug-resistant Plasmodium falciparum. Science 244: 1184–1186

15. Cowman AF, Galatis D, Thompson JK (1994) Selection of mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 and cross-resistance to halofantrine and quinine. Proceedings of the National Academy of Sciences 91: 1143–1147

16. Gassis S, Rathod P (1996) Frequency of drug resistance in Plasmodium falciparum: a nonsynergistic combination of 5-fluoroorotate and atovaquone suppresses in vitro resistance. Antimicrob Agents Chemother 40: 914–919

17. Rathod PK, McErlean T, Lee P-C (1997) Variations in frequencies of drug resistance in Plasmodium falciparum. Proceedings of the National Academy of Sciences 94: 9389–9393

18. Su X-z, Kirkman LA, Fujioka H, Wellems TE (1997) Complex Polymorphisms in an ~330 kDa Protein Are Linked to Chloroquine-Resistant P. falciparum in Southeast Asia and Africa. Cell 91: 593–603

19. Eastman RT, White J, Hucke O, Bauer K, Yokoyama K, et al. (2005) Resistance to a protein farnesyltransferase inhibitor in Plasmodium falciparum. J Biol Chem 280: 13554–13559

20. Singh A, Rosenthal PJ (2004) Selection of cysteine protease inhibitor-resistant malaria parasites is accompanied by amplification of falcipain genes and alteration in inhibitor transport. J Biol Chem 279: 35236–35241

21. Dharia N, Sidhu A, Cassera M, Westenberger S, Bopp S, et al. (2009) Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in Plasmodium falciparum. Genome Biology 10: R21.

22. Rottmann M, McNamara C, Yeung BK, Lee MC, Zou B, et al. (2010) Spiroindolones, a potent compound class for the treatment of malaria. Science 329: 1175–1180

23. Eastman RT, Dharia NV, Winzeler EA, Fidock DA (2011) Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured Plasmodium falciparum parasites. Antimicrob Agents Chemother 55: 3908–3916

24. Cui L, Wang Z, Miao J, Miao M, Chandra R, et al. (2012) Mechanisms of in vitro resistance to dihydroartemisinin in Plasmodium falciparum. Mol Microbiol 86: 111–128

25. Ganesan K, Ponmee N, Jiang L, Fowble JW, White J, et al. (2008) A Genetically hard-wired metabolic transcriptome in Plasmodium falciparum fails to mount protective responses to lethal antifolates. PLoS Pathog 4: e1000214.

26. Eastman RT, White J, Hucke O, Yokoyama K, Verlinde CL, et al. (2007) Resistance mutations at the lipid substrate binding site of Plasmodium falciparum protein farnesyltransferase. Mol Biochem Parasitol 152: 66–71

27. Freeman DL, Ponmee N, Guler JL, White J, Gujjar R, et al. . (2009) Target gene duplication in ARMD plasmodium falciparum acquiring drug resistance. Woods Hole Molecular Parasitology Meeting XX. September 13–17 ed.

28. Istvan ES, Dharia NV, Bopp SE, Gluzman I, Winzeler EA, et al. (2011) Validation of isoleucine utilization targets in Plasmodium falciparum. Proc Natl Acad Sci U S A 108: 1627–1632

29. Phillips MA, Gujjar R, Malmquist NA, White J, El Mazouni F, et al. (2008) Triazolopyrimidine-based dihydroorotate dehydrogenase inhibitors with potent and selective activity against the malaria parasite Plasmodium falciparum. Journal of Medicinal Chemistry 51: 3649–3653

30. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, et al. (2009) PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res 37: D539–543

31. Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res 26: 4056–4062

32. Triglia T, Foote SJ, Kemp DJ, Cowman AF (1991) Amplification of the multidrug resistance gene pfmdr1 in Plasmodium falciparum has arisen as multiple independent events. Mol Cell Biol 11: 5244–5250

33. Nair S, Nash D, Sudimack D, Jaidee A, Barends M, et al. (2007) Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol Biol Evol 24: 562–573

34. Barnes DA, Foote SJ, Galatis D, Kemp DJ, Cowman AF (1992) Selection for high-level chloroquine resistance results in deamplification of the pfmdr1 gene and increased sensitivity to mefloquine in Plasmodium falciparum. EMBO J 11: 3067–3075

35. Thaithong S, Ranford-Cartwright LC, Siripoon N, Harnyuttanakorn P, Kanchanakhan NS, et al. (2001) Plasmodium falciparum: gene mutations and amplification of dihydrofolate reductase genes in parasites grown in vitro in presence of pyrimethamine. Experimental Parasitology 98: 59–70

36. Andersson DI, Slechta ES, Roth JR (1998) Evidence that gene amplification underlies adaptive mutability of the bacterial lac operon. Science 282: 1133–1135

37. Preechapornkul P, Imwong M, Chotivanich K, Pongtavornpinyo W, Dondorp AM, et al. (2009) Plasmodium falciparum pfmdr1 amplification, mefloquine resistance, and parasite fitness. Antimicrob Agents Chemother 53: 1509–1515

38. Nair S, Miller B, Barends M, Jaidee A, Patel J, et al. (2008) Adaptive copy number evolution in malaria parasites. PLoS Genet 4: e1000243.

39. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498–511

40. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, et al. (2007) Vivax malaria: neglected and not benign. Am J Trop Med Hyg 77: 79–87

41. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, et al. (2008) Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature 455: 757–763

42. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, Leishmania major. Science 309: 436–442

43. Beverley SM (1991) Gene amplification in Leishmania. Annu Rev Microbiol 45: 417–444

44. Koo HS, Wu HM, Crothers DM (1986) DNA bending at adenine. thymine tracts. Nature 320: 501–506

45. Hile SE, Eckert KA (2008) DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellite sequences. Nucleic Acids Res 36: 688–696

46. Schultes NP, Szostak JW (1991) A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in Saccharomyces cerevisiae. Mol Cell Biol 11: 322–328

47. Ira G, Haber JE (2002) Characterization of RAD51-independent break-induced replication that acts preferentially with short homologous sequences. Mol Cell Biol 22: 6384–6392

48. Bzymek M, Lovett ST (2001) Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc Natl Acad Sci U S A 98: 8319–8325

49. Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10: 551–564

50. Slack A, Thornton PC, Magner DB, Rosenberg SM, Hastings PJ (2006) On the mechanism of gene amplification induced under stress in Escherichia coli. PLoS Genet 2: e48.

51. Lee JA, Carvalho CM, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131: 1235–1247

52. Payen C, Koszul R, Dujon B, Fischer G (2008) Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. PLoS Genet 4: e1000175.

53. Tyagi S, Sharma M, Das A (2011) Comparative genomic analysis of simple sequence repeats in three Plasmodium species. Parasitol Res 108: 451–458

54. Lovett ST, Hurley RL, Sutera VA Jr, Aubuchon RH, Lebedeva MA (2002) Crossing over between regions of limited homology in Escherichia coli. RecA-dependent and RecA-independent pathways. Genetics 160: 851–859

55. Deng X, Gujjar R, El Mazouni F, Kaminsky W, Malmquist NA, et al. (2009) Structural plasticity of malaria dihydroorotate dehydrogenase allows selective binding of diverse chemical scaffolds. Journal of Biological Chemistry 284: 26999–27009

56. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, et al. (2006) A systematic map of genetic variation in Plasmodium falciparum. PLoS Pathog 2: e57.

57. Sander AF, Salanti A, Lavstsen T, Nielsen MA, Theander TG, et al. Positive selection of Plasmodium falciparum parasites with multiple var2csa-type PfEMP1 genes during the course of infection in pregnant women. J Infect Dis 203: 1679–1685

58. Koita OA, Doumbo OK, Ouattara A, Tall LK, Konare A, et al. (2012) False-negative rapid diagnostic tests for malaria and deletion of the histidine-rich repeat region of the hrp2 gene. Am J Trop Med Hyg 86: 194–198

59. Haynes JD, Diggs CL, Hines FA, Desjardins RE (1976) Culture of human malaria parasites Plasmodium falciparum. Nature 263: 767–769

60. Trager W, Jensen JB (1976) Human malaria parasites in continuous culture. Science 193: 673–675

61. Rathod PK, Leffers NP, Young RD (1992) Molecular targets of 5-fluoroorotate in the human malaria parasite, Plasmodium falciparum. Antimicrob Agents Chemother 36: 704–711

62. Hu G, Llinas M, Li J, Preiser P, Bozdech Z (2007) Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. BMC Bioinformatics 8: 350.

63. Gonzales JM, Patel JJ, Ponmee N, Jiang L, Tan A, et al. (2008) Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. PLoS Biol 6: e238.

64. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 1: e5.

65. Tan JC, Patel JJ, Tan A, Blain JC, Albert TJ, et al. (2009) Optimizing comparative genomic hybridization probes for genotyping and SNP detection in Plasmodium falciparum. Genomics 93: 543–550

66. Samarakoon U, Gonzales JM, Patel JJ, Tan A, Checkley L, et al. (2011) The landscape of inherited and de novo copy number variants in a Plasmodium falciparum genetic cross. BMC Genomics 12: 457.

67. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucl Acids Res 29: e45.

68. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

69. Biggs BA, Kemp DJ, Brown GV (1989) Subtelomeric chromosome deletions in field isolates of Plasmodium falciparum and their relationship to loss of cytoadherence in vitro. Proc Natl Acad Sci U S A 86: 2428–2432

70. Kemp DJ, Thompson J, Barnes DA, Triglia T, Karamalis F, et al. (1992) A chromosome 9 deletion in Plasmodium falciparum results in loss of cytoadherence. Mem Inst Oswaldo Cruz 87 Suppl 3: 85–89

71. Anderson TJ, Patel J, Ferdig MT (2009) Gene copy number and malaria biology. Trends Parasitol 25: 336–343

72. Jiang L, Lee P-C, White J, Rathod PK (2000) Potent and selective activity of a combination of thymidine and 1843U89, a folate-based thymidylate synthase inhibitor, against Plasmodium falciparum. Antimicrob Agents Chemother 44: 1047–1050

73. Rathod PK, Khosla M, Gassis S, Young RD, Lutz C (1994) Selection and characterization of 5-fluoroorotate-resistant Plasmodium falciparum. Antimicrob Agents Chemother 38: 2871–2876

74. Young RD, Rathod PK (1993) Clonal viability measurements on Plasmodium falciparum to assess in vitro schizonticidal activity of leupeptin, chloroquine, and 5-fluoroorotate. Antimicrob Agents Chemother 37: 1102–1107

75. Cui L, Su XZ (2009) Discovery, mechanisms of action and combination therapy of artemisinin. Expert Rev Anti Infect Ther 7: 999–1013

Chapter 3: Deep sequencing of increased artemisinin resistant laboratory strains reveal a novel genomic amplification on chromosome 10

This chapter is a summary of work done by:

Vida Ahyong, Matthew Tucker, Katherine Sorber, Dennis Kyle, and Joseph DeRisi

Author Contributions:

Vida Ahyong did the library preparations and deep sequencing analyses. Matthew Tucker provided the genomic DNA. Katherine Sorber performed the preliminary sequencing and qPCR experiments. Dennis Kyle and Joseph DeRisi conceived and supervised the project.

Introduction

Malaria continues to take a devastating toll on global health with approximately 225 million cases worldwide resulting in >500,000 deaths reported in 2013[1]. Efforts to eradicate malaria are hindered by several factors including drug resistance to anti-malarial drugs. Currently, the World Health Organization (WHO) recommends artemisinin combination therapy (ACTs) as a first line of treatment in endemic areas[2]. Several reports in the last decade describe artemisinin resistance on the Thai-Cambodian border by longer parasite clearance times in both monotherapy and combination therapy, elevated IC50s, increased gametocyte carriage, and longer clearance times since ACTs were introduced in 1995[3][4][5]. This phenomenon is especially important in light of the geographical location of these parasites on the Thai-

Cambodian border, a historical source of novel anti-malarial drug resistance. SNPs and amplifications observed in PfMDR1, the *P. falciparum* multi-drug resistance protein 1, indicated that this gene may have relevance to artemisinin resistance but Pfmdr1 since been described as a modulator of resistance and not the causal resistance gene[6][7].

Previous studies of artemisinin resistance in field isolates as well as *in vitro* selected parasites have identified global transcriptome stalls in ring stage parasites that may contribute to this drug resistant phenotype[8]. In this study, we use whole genome sequencing data to compare *in vitro* selected artemisinin resistant parasite lines to sensitive parental lines to detect genetic changes consistently associated with a high level of resistance to artemisinins *in vitro[9].* By comparing parental strains (D6s and W2s) that remain sensitive to artemisinin and drug selected strains (D6r and W2r), we uncovered a genome amplification present in the artemisinin resistant strains that should be tested as a candidate resistant locus in field isolates. This region is comprised of a 30 kb region in chromosome 10 that is amplified in both D6r and W2r but is not present in either parental strain. Genome wide SNP analysis revealed that there are no common SNPs between the two resistant strains leading us to speculate that the chromosome 10 amplification may contain a gene product that can influence the increased resistance to artemisinin in our *in vitro* model.

Materials and Methods

Genome DNA Extraction

Extraction of genomic DNA was prepared by two sorbitol synchronizations to achieve >95% ring stage parasites and harvested by lysing red blood cells in a final concentration of 0.1% saponin in 1xPBS. Genomic DNA was extracted by adding a final concentration of 0.1%

sarkosyl and 0.25U of Proteinase K and incubating overnight at 37°C. Two rounds of phenol-chloroform extraction in 5Prime Phase lock light tubes (5Prime Inc., Gaithersburg, MD) were performed, followed by the addition of RNAse I (Ambion, Austin, TX) at 37°C for 1 hour. An additional phenol-chloroform and chloroform only extraction was followed by precipitation with 3M NaOAc, and ethanol precipitation. Final gDNA pellets were resuspended in 50ul of water and concentration determined on a Nanodrop spectrophotometer (Thermo Scientific,Wilmington, DE).

Paired End Library Preparation

Samples for sequencing the artemisinin sensitive (D6s and W2s) and resistant strains (D6R and W2R) were prepared using the Illumina paired end sample kit (Illumina, Hayward, CA). 2-5 ug of genomic DNA was nebulized at 32 psi for 6 minutes and purified using Zymo-Spin Columns (Zymo Research, Orange, CA) or Qiagen MinElute Columns (Qiagen Inc., Valencia, CA). End repair, ligation of adapters, and PCR enrichment were performed as described in the Illumina Paired End Sample Kit protocol and purified and concentrated by spin columns after each step. Ligation products were purified on a 2% TAE Agarose gel, extracting a range of 250-350 bp fragment using Invitrogen's PureLink Gel Extraction Kit (Invitrogen, Carlsbad, CA). Validation of sequencing libraries was performed by TOPO cloning (Invitrogen, Carlsbad, CA) 4 ul of the PCR enriched library and sequencing of 30 clones. A final library concentration of 8 pM was loaded into a V4 flow cell and clusters were generated using the Illumina Cluster Generation Kit and sequenced on the Illumina GAii (Illumina, Hayward, CA).

Downstream Sequencing Analysis

Reads were aligned to the *Plasmodium falciparum* genome (PlasmoDB version 7.1) using the Bowtie short read aligner[10]. Reads were filtered to remove reads mapped to the human genome before aligning to the Pf genome. Since the published genome is based on the 3D7 strain, we allowed for one mismatch within a read to accommodate possible strain specific SNP differences with D6 and W2. Any reads with more than 2 mismatches were not further analyzed. We also required that reads could be mapped to a unique location within the genome. We calculated coverage per base pair throughout the entire genome and created WIG files to view the sequenced genomes on the UCSC genome browser.

QPCR Validation of Amplified Region

Relative copy number was validated by qPCR using the Lightcycler 480 Sybr Green I Master Mix(Roche Applied Science, Indianapolis, IN) from the region surrounding and within the amplified region with 0.5 uM primer and 1 ng of genomic DNA. Cycling was performed and analyzed on the Roche Lightcycler 480 (Roche Applied Science, Indianapolis, IN). Cycling conditions were 95 °C for 10 min, followed by 40 cycles of 95 °C for 30 sec, 52 °C for 45 sec, 65 °C for 1 min, and 68 °C for 10 sec. Fluorescence was read following each cycle and a final extension was performed at 65 °C for 7 min before melting curve analysis was performed (65 °C to 95°C with a 5 sec hold for every 0.5 °C followed by a fluorescence read). Only reactions with one clean peak from melting curve analysis were analyzed. Genomic DNA from ring stage 3D7 Oxford parasites served as the control sample, while PFL2510w (chitinase) served as the reference gene. See Table 4 for list of qPCR primers used in this study.

Results

Sequencing of Artemisinin Sensitive and Resistant Strains

The clones of lab strains D6 and W2 were selected for artemisinin resistance at 2400 ng/ul and 200 ng/ul, respectively as previously described in Tucker et al[9]. Genomic DNA was isolated and purified from D6 and W2 sensitive and resistant strains and prepared for 65 bp length Illumina sequencing. The resulting libraries contained 7.1 million reads for W2s, 11.5 million reads for W2r, 12.9 million reads for D6s, and 55.7 million reads (Table 1).

Genome Wide SNP Analysis

A custom script for whole genome SNP analysis was used to identify SNPs between the sensitive and resistant strains using SAM files generated from bowtie alignment files. We reasoned that if a SNP were responsible for conferring artemisinin resistance, it would most likely reside within an exon of a non-antigenic variation, protein-coding gene. Therefore, we looked for exon positions in the genome with at least five reads of coverage for each strain where the dominant bp differed between the parental and resistant strains. Thresholds were set to contain over 90% SNP agreement for both the sensitive and resistant allele. This list was further narrowed by assuming that a causative SNP was also most likely to cause a non-synonymous rather than synonymous amino acid change in the resulting protein (Table 2 & Table 3). For W2 we found 6 SNPs and for D6 we found 5 SNPs that passed our analysis requirements.

Detection and Verification of Genome Amplification

Because of characteristically uneven coverage throughout the *P. falciparum* genome due to high A/T, low complexity regions that are difficult to map, programmatic determination of genome amplifications required smoothing of the data. Based on empirical trial-and-error, mean coverage of 7500 bp windows were plotted for the entire genome (Figure 1). We identified a region in chromosome 10 in which the D6 resistant strain showed approximately a contiguous 74 kb of 2-fold amplification in read coverage, absent in the parental strain. This region spanned from PF10_0279 – PF10_0299 and copy number was determined by the ratio of average coverage per base pair of the resistant to the sensitive strain (Figure 2). Similarly, the W2 resistant strain showed a contiguous 30 kb region of 3-fold amplification contained within the boundaries of the D6 amplification. This amplified region spanned PF10_0288 to PF10_0297.

To verify these amplifications and more accurately determine the edges of the chromosome 10 amplification, qPCR primers were designed to several genes both within and near the presumed edges of the W2 strains (Figure 3). Our qPCR results showed the lack of copy number differences at gene PF10_0287 which is just to the 5' end of the amplicon, and at gene PF10_0298 which is just to the 3' end of the amplicon. Whereas seven genes internal to the amplicon showed a two to five-fold difference in copy number by qPCR.

In addition to the chromosome 10 amplification, we also found a 100 kb region of amplification in chromosome 5 in the W2 resistant strain but not in the D6 resistant strain. The region includes the PfMDR1 gene and could act to further modulate the resistance to artemisinin in conjunction with the causative resistant gene.

Finally, we asked whether single nucleotide polymorphisms could contribute to the resistance phenotype. We developed a custom single nucleotide polymorphism script that takes in aligned deep sequencing reads from Bowtie and calculates the nucleotide frequencies of every single position of the genome. By filtering only for positions that have a mutation in the resistant strain with ample coverage, purity (See Materials and Methods) and a non-synonymous change within an exon results in a list of SNPs that differentiate the sensitive strain from the resistant strain. Though we found several SNPs in the W2 and D6 strains, we found no single SNP or gene with mutations that was present in both W2 and D6 (Table 2 & 3). Despite the lack of overlap between the two SNP gene lists, there is still a possibility that these mutations could help modulate the individual strains resistance profile.

Discussion

Here we have established new genetic leads in search for the element responsible for artemisinin resistance in *P. falciparum* by sequencing two independent *in vitro* selected parasite lines, D6 and W2. Classically, pathogen drug resistance can arise by mutation of the drug's target (loss of function), or enhanced metabolism or efflux of the drug (gain of function). These mechanisms are most likely to arise from changes in the amino acid composition of proteins, non-synonymous mutations. Our SNP analysis reveals a small subset of genes with non-synonymous mutations gained during the selection process. We expect that a causative resistant SNP or gene would be present in both the resistant strains when compared to their respective artemisinin-sensitive parent strain. However our deep sequencing results indicate that there are no overlapping SNPs between the pairs of artemisinin sensitive and resistant strains, leading us to suspect that a SNP is not responsible for drug resistance. Thus, we manually scanned the entire genomic profile for alternate genetic changes that correlate with resistance, including insertions, deletions, and amplifications.

From our deep sequencing data, we are able to detect genomic structural changes based on mapping reads across the complete *P. falciparum* genome and identifying gaps or increased coverage levels that indicate structural rearrangements. Because of intrinsic differences between our W2 and D6 strains and the published 3D7 genome, we expected to see numerous coverage level variations in our genome browser. However with the same reasoning as our SNP analysis, we expect that if a structural variant is responsible for a common resistance mechanism, both the artemisinin resistant lines would show a significant difference in coverage levels that would not be present in the sensitive parental lines. We detected a novel genome amplification in chromosome 10 that is our most intriguing candidate for drug resistance since both resistant strains contain at least a 30 kb amplified region from PF10_0288 to PF10_0297. Of the 10 genes covering this region, four have annotated putative gene functions and six are conserved genes with unknown functions[table of genes]. To truly establish a causal relationship between the chromosome 10 amplification and resistance to artemisinins, more research on these genes are required, either through biochemical characterization of these genes in the presence of artemisinin or detecting increased resistance through transfections of individual genes into wild type parasites. Surveys of these genetic leads, both SNPs and amplifications, should be tested in clinical isolates from areas where resistance is recognized (Thai-Cambodian border) to ascertain whether similar mutations are associated with resistance in the clinic.

References

1. World Health Organization, Global Malaria Programme, World Health Organization. World Malaria Report 2014. 2014.

2. World Health Organization. Guidelines for the treatment of malaria. Geneva: World Health Organization; 2015.

3. Dondorp AM, Yeung S, White L, Nguon C, Day NPJ, Socheat D, Von Seidlein L. Artemisinin resistance: current status and scenarios for containment. Nat Rev Microbiol. 2010 Apr;8(4):272–280.

4. Noedl H, Socheat D, Satimai W. Artemisinin-resistant malaria in Asia. N Engl J Med. 2009 Jul 30;361(5):540–541.

5. Carrara, V.I. et al. Changes in the Treatment Responses to Artesunate-Mefloquine on the Northwestern Border of Thailand during 13 Years of Continuous Deployment. *PLoS ONE* **4**, e4551 (2009).

6. Duraisingh MT, Cowman AF. Contribution of the pfmdr1 gene to antimalarial drug-resistance. Acta Trop. 2005 Jun;94(3):181–190.

7. Duraisingh MT, Refour P. Multiple drug resistance genes in malaria -- from epistasis to epidemiology. Mol Microbiol. 2005 Aug;57(4):874–877.

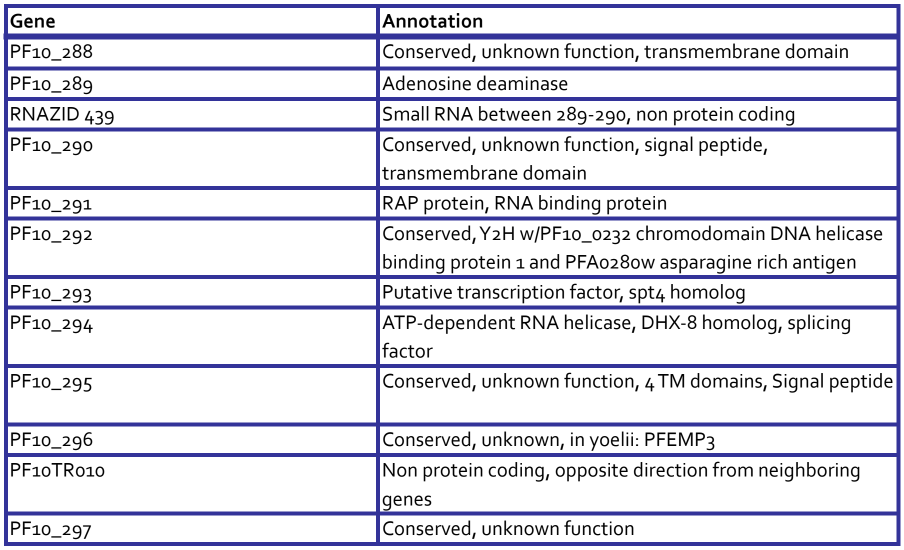
8. Mok S, Imwong M, Mackinnon MJ, Sim J, Ramadoss R, Yi P, Mayxay M, Chotivanich K, Liong K-Y, Russell B, Socheat D, Newton PN, Day NPJ, White NJ, Preiser PR, Nosten F, Dondorp AM, Bozdech Z. Artemisinin resistance in Plasmodium falciparum is

associated with an altered temporal pattern of transcription. BMC Genomics. 2011;12:391.

9. Tucker MS, Mutka T, Sparks K, Patel J, Kyle DE. Phenotypic and genotypic analysis of in vitro-selected artemisinin-resistant progeny of Plasmodium falciparum. Antimicrob Agents Chemother. 2012 Jan;56(1):302–314.

10. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

Figure 1: Genes within the chromosome 10 amplified region and their annotated functions.



Gene	Annotation
PF10_288	Conserved, unknown function, transmembrane domain
PF10_289	Adenosine deaminase
RNAZID 439	Small RNA between 289-290, non protein coding
PF10_290	Conserved, unknown function, signal peptide, transmembrane domain
PF10_291	RAP protein, RNA binding protein
PF10_292	Conserved, Y2H w/PF10_0232 chromodomain DNA helicase binding protein 1 and PFA0280w asparagine rich antigen
PF10_293	Putative transcription factor, spt4 homolog
PF10_294	ATP-dependent RNA helicase, DHX-8 homolog, splicing factor
PF10_295	Conserved, unknown function, 4 TM domains, Signal peptide
PF10_296	Conserved, unknown, in yoelii: PFEMP3
PF10TR010	Non protein coding, opposite direction from neighboring genes
PF10_297	Conserved, unknown function

Figure 2: A sliding scale window of 7500bp of the chromosome 10 amplification shows an increase in the copy number of the region for both the D6 and W2 resistance strain of 1n to 3n for D6 and 2n to 3n for W2.

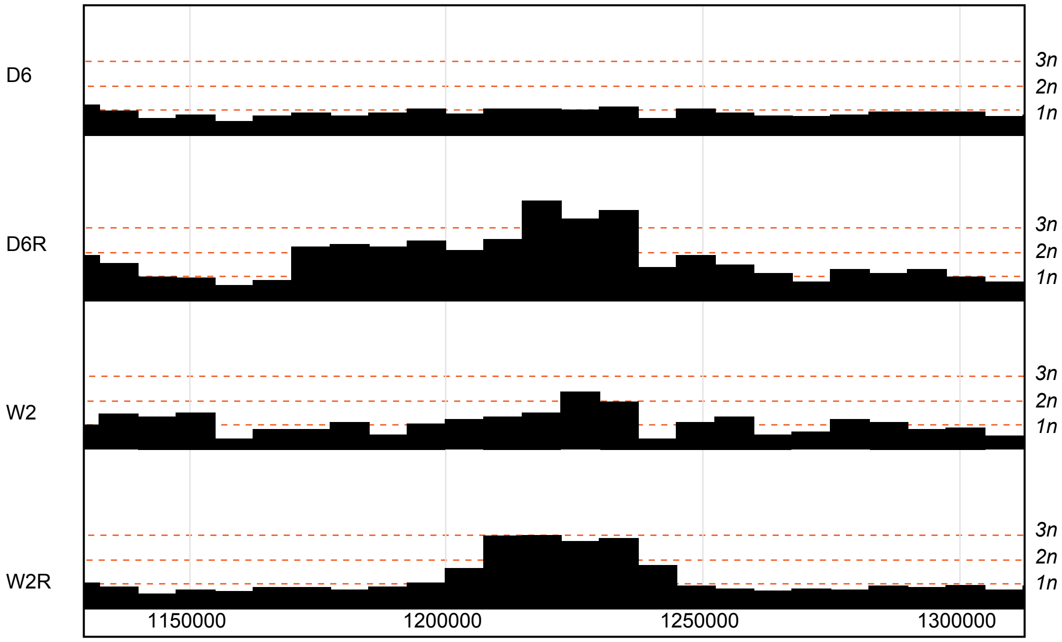


Figure 3: Sequencing coverage and qPCR validation. The upper panel shows the sequencing coverage of the W2 resistant clone with amplicon borders between genes PF10_0288 to PF10_0297 (in orange). qPCR validation across the unamplified and amplified region between the W2 sensitive (white bars) and the W2 resistant (green bars) shows the relative qPCR copy numbers for selected genes.

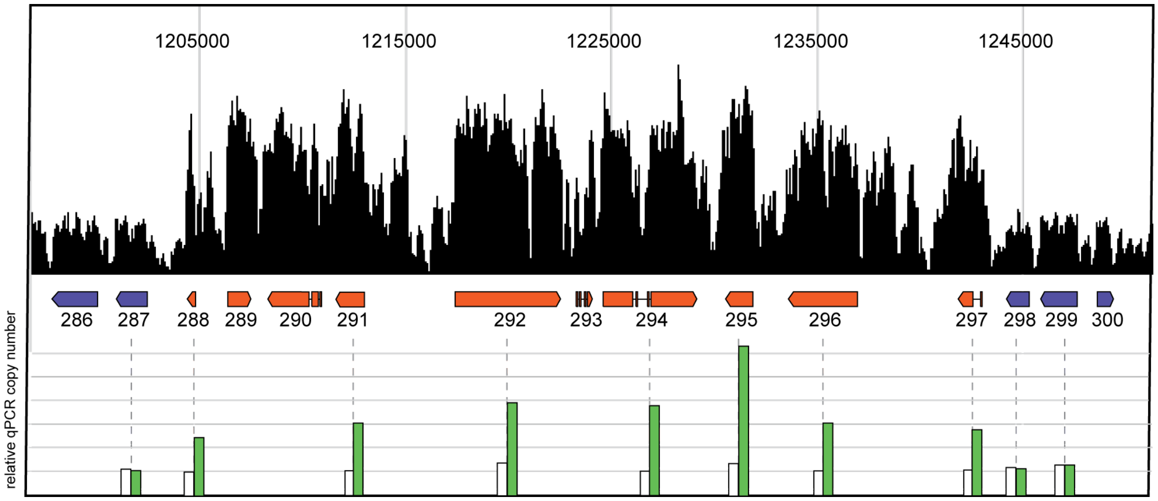


Table 1: Sequencing statistics of the four libraries in this study.

Strain	# aligned reads	Average genome coverage	Single or Paired end?
D6	12.93 million	~50x	Single
D6 Resistant*	55.75 million	~150x	Paired (aligned using only single reads)
W2 **	7.51 million	~20x	single
W2 Resistant	11.53 million	~35x	single

*Uneven coverage, exon specific, due to fragmentase

** Our W2, not from the Kyle lab

Table 2: SNP profile for W2 resistant strains.

Chromosome	SNP	Gene	Description
chr5	Gly -> Cys	PFE1150w	PFMDR1
chr3	Glu -> Asp	PFC0135c	Putative exportin, crm1 homolog
chr3	Asp -> Tyr	PFC0625w	Conserved Plasmodium membrane protein
chr7	Gln -> His	PF07_0126	Transcription factor with AP2 domain, putative
chr9	Asp -> Tyr	PFI0850w	Conserved Plasmodium protein
chr14	Leu -> Val	PF14_0139	Conserved Plasmodium protein

Table 3: SNP profile of D6 artemisinin resistant strains.

Chromosome	SNP	Gene	Description
chr3	Asn -> Ile	PFC0320w	Conserved Plasmodium protein
chr4	Arg -> Gly	PFD0900w	Conserved Plasmodium protein
chr8	Stop -> Glu	MAL8P1.212	RESA like pseudogene
chr8	Arg -> Ile	MAL8P1.212	RESA like pseudogene
chr4	Arg -> Lys	PFD0900w	Conserved Plasmodium protein

Table 4: qPCR primers used in this study to detect copy number variations.

Primer Name	5’ -> 3’
287F	GAAATTATGCACAAGGCTAGTTCC
287R	GGAACTAGCCTTGTGCATAATTTC
288F	CGTGCTTATTATTACTATCAGG
288R	GGTATATGATAAGAGGTATGTTC
291F	ACGCTGCTTAGACCCAGAGA
291R	AATGTTGCCGCTTTTTGTATG
292F	CGCACACAAACACACGTACA
292R	CCGAATCTTCGTTACCTGGA
294F	CGTCCTGAATATCCACCTGAA
294R	TCACACTCTGCATTTCTGACG
295F	CCCTCAACAATCAAGGCAAT
295R	CATGTCCCCAAATTTCATCC
296F	AACATTTTCACGCGACTTCC
296R	TGTGCGTTTTGCTCCAATAA
297F	CCTATATTATTTTCGTGGTTATACC
297R	GTAAAAATAAATTATGCATAAAG
298F	GGGAAAGCGATAAATATAATC
298R	GATTATATTTATCGCTTTCCC
299F	TTCATTGCATCCTTGATTGG
299R	AATGCACCCTCACCAGGATA
chitinase F	TGTTTCCTTCAACCCCTTTT
chitinase R	TAATCCAAACCCGTCTGCTC

Chapter 4: Genome-wide Regulatory Dynamics of Translation in *the Plasmodium falciparum* Asexual Blood Stages

This chapter is a reprint from the following reference:

Florence Caro†, Vida Ahyong†, Miguel Betegon, Joseph L. DeRisi. Genome-wide Regulatory Dynamics of Translation in the *Plasmodium falciparum* Asexual Blood Stages. eLife 2014;3:e04106

† These authors contributed equally

Author Contributions

FC, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article

VA, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article

MB, Conception and design, Analysis and interpretation of data

JLDR, Conception and design, Analysis and interpretation of data, Drafting or revising the article

Abstract

The characterization of the transcriptome and proteome of *Plasmodium falciparum*, has been a tremendous resource for the understanding of the molecular physiology of this parasite. However, the translational dynamics that link steady-state mRNA with protein levels are not well understood. Here, we bridge this disconnect by measuring genome-wide translation using ribosome profiling, through five stages of the *P. falciparum* blood phase developmental cycle. Our findings show that transcription and translation are tightly coupled, with overt translational control occurring for less than 10% of the transcriptome. Translationally regulated genes are predominantly associated with merozoite egress functions. We systematically define mRNA 5' leader sequences, and 3'UTRs, as well as antisense transcripts, along with ribosome occupancy for each, and establish that accumulation of ribosomes on 5' leaders is a common transcript feature. This work represents the highest resolution and broadest portrait of gene expression and translation to date for this medically important parasite.

Introduction

The transcriptome of the intraerythrocytic developmental cycle (IDC) is characterized by a continuous cascade wherein the expression of the majority of genes are maximally induced once per cycle and their timing correlates well with the timing for the respective protein's biological function (Bozdech et al. 2003). The apparent lack of dynamic transcriptional regulation suggested that complementary post-transcriptional mechanisms could play an important role in the regulation of parasite gene expression (Hughes et al. 2010). This is a reasonable assumption, given that global or gene-specific translational regulation of gene expression is a mechanism that allows fast adaptations during drastic changes in environmental conditions as well as during rapid transitions in developmental programs. Indeed a few examples of translational control in *Plasmodium* have been reported. In sporozoites present in the mosquito salivary gland, phosphorylation of the eukaryotic translation initiation factor eIF2α by the kinase IK2, inhibits translation and causes accumulation of mRNAs into granules. Translational repression is alleviated by eIF2α phosphatase during the transition into the mammalian host, allowing parasites to transform into the liver stages (Zhang et al. 2010). Similarly, PK4 kinase activity leads to the reduction of global protein synthesis through phosphorylation of eIF2α in schizonts and gametocytes and is essential for the completion of the parasite's erythrocytic cycle (Zhang et al. 2012). Gene-specific translational regulation has also been observed in *P. falciparum* and is mediated by cis-acting sequences in combination with RNA-binding proteins. For example, dihydrofolate reductase-thymidylate synthase (DHFR-TS) binds within the coding region of its own cognate mRNA to repress translation (Zhang & Rathod 2002) and antifolate treatment has been shown to relieve this repressive effect without alteration of mRNA levels (Nirmalan et al. 2004). In *Plasmodium berghei,* storage of translationally repressed mRNAs prior to fertilization is mediated by mRNA binding via the RNA helicase DOZI and the Sm-like factor CITH (Mair et al. 2006; Mair et al. 2010). Upstream open reading frames (uORFs) found on 5' UTRs of transcripts have been reported to regulate the translation of specific genes (Morris & Geballe 2000). In *P. falciparum*, the only uORF described and functionally characterized to date is a 120 codon region upstream of the *var2csa* (PFL0030c) coding region, a unique variant of the surface antigen PfEMP1 that mediates adhesion to placenta in pregnant women (Amulic et al. 2009). In this case, translation of the uORF modulates repression of var2csa translation. Aside from these examples, the extent to which global and gene-specific translational control operates in *P. falciparum* during the IDC remains sparse.

Since the *Plasmodium falciparum* genome was fully sequenced (Gardner et al. 2002), several large-scale studies have provided detailed insights into the expression of genes and proteins across the parasite's life cycle. Parallel mass spectrometry-based proteomics and genome-wide expression profiling revealed differences between mRNA abundance and the accumulation of the corresponding protein, supporting the notion that post-transcriptional regulation of gene expression is at play in this parasite (Nirmalan et al. 2004; Le Roch et al. 2004; Foth et al. 2011). These methods, however, are limited in their ability to measure low abundance proteins and do not capture the underlying relationship between transcriptional activity and translational efficiency. More recently, polysome profiling was used to monitor discrepancies between polysome-associated and steady state mRNAs in 30% of the *P. falciparum* blood stage transcriptome (Evelien M Bunnik et al. 2013), however this approach does not reveal the precise localization of the ribosomes, and thus can not be used to accurately assess the translational efficiency of a given mRNA (Ingolia 2014).

Here, we adapted the ribosome profiling technique (Ingolia et al. 2009) to describe the translational dynamics of the *P. falciparum* asexual blood stage transcriptome. We simultaneously evaluate mRNA abundance, gene structure, ribosome positioning, and translational efficiency for genes expressed through five stages of the IDC. We demonstrate that the data are highly reproducible, and we find that the translational efficiency of the majority of mRNAs expressed follows a narrow distribution, exhibiting a tight coupling between transcription and translation. Only 10% of the genes expressed deviate from this trend and are translationally up- or down-regulated. We found a surprising amount of ribosome density associated with 5’ leaders of transcripts particularly in genes with functions associated to merozoite egress and invasion. Overall, the precision and depth of the dataset presented herein add significantly to our understanding of *P. falciparum* gene expression by linking transcriptional and translational dynamics throughout the blood stages.

Results

Overview of ribosome profiling in *P. falciparum* asexual blood stages

To create whole-genome, high-resolution profiles of mRNA abundance and translation during *in vivo* blood stage development of *P. falciparum,* we adapted the ribosome profiling technique described by Ingolia *et al*. (Ingolia et al. 2009). Ribosome profiling is based on the deep sequencing of ribosome-protected mRNA fragments obtained by nuclease-digestion of polysomes, cycloheximide-arrested ribosomes bound to mRNA. These fragments represent the exact location of the ribosome at the moment the sample was harvested. Five stages representative of the 48-hour IDC of *P. falciparum* were harvested for both mRNA and polysome isolation; ring, early trophozoite, late trophozoite, schizont stages and purified merozoites. To assess the reproducibility of the data, we harvested independent biological replicates of each stage. Polysomes were isolated in the presence of the translation elongation inhibitor cycloheximide, then nuclease digested to produce monosomes, and sedimented by centrifugation on a sucrose gradient (Figure 1-figure supplement 1). To minimize isolation of RNA fragments bound by proteins other than 80S ribosomes, RNA was extracted only from the fractions of the sucrose gradient containing the monosome peak. The resulting ~30 nt fragments of RNA, corresponding to ribosome footprints, were processed into strand specific deep sequencing libraries in parallel with the mRNA samples, fragmented to ~30 nt for consistency. Despite the unusually high AT content of the *P. falciparum* genome, over 92% of all 30 nt sequenced reads, derived from coding sequences (CDSs), mapped uniquely to the genome

(Figure 1-Source Data 1 and Figure 1-figure supplement 2).

To quantitatively obtain mRNA abundance and ribosome footprint density measures, we calculated rpkMs (reads per kilobase of exon model per million reads mapped, as in (Mortazavi et al. 2008)) for each gene. We established the minimum number of mRNA reads sequenced per coding region (rM; reads per million reads mapped) required to confidently include genes in downstream analyses, to be ≥32 rM (Materials and Methods, Figure 1-figure supplement 3). Using this conservative threshold, 3,605 genes qualified for further analysis. Between biological replicates, Pearson correlation values were consistently high, ranging from r =0.94 to r =0.99 (Figure 2A), highlighting the quality and reproducibility of our data. In addition, we compared the RNA-seq transcriptome of the five stages sampled to our previously published transcriptome dataset, originally generated using long oligo microarrays (Bozdech et al. 2003). The RNA-Seq transcription profiles of the set of genes shared by the two datasets (n = 1829) were highly correlated (average r = 0.7) to the corresponding 11, 21, 31, 45, and 2 hours post merozoite invasion time points of the microarray dataset, despite the use of different methodologies (microarray vs. RNA-seq), and the use of different *P. falciparum* strains (HB3 vs. W2, respectively). Because of the higher sensitivity of RNA-seq we were able to accommodate an additional 743 genes into the cascade-like transcriptome extending it to a total of 3110 genes (Figure 2B, Figure 2-Source data 1). The remaining 495 genes in our RNA-seq dataset lacked sufficient variation over the five time points for inclusion within the phaseogram. These genes, referred to as non-phasic genes, are nevertheless included in all analyses.

Gene expression and translation are tightly coupled during the *P. falciparum* IDC

While RNA-Seq reveals the abundance and architecture of individual mRNAs, ribosome profiling provides a complementary and quantitative measure of mRNA translation. Ribosome occupancy along the CDS results in a profile that indicates the timing and magnitude of translation of a given mRNA, thus quantitatively delineating regions of each mRNA molecule that is actually bound by 80S ribosomes (Ingolia et al. 2009). To inspect translation on a genome-wide scale, ribosome density values of each gene expressed in the dataset were organized in the same order as the transcriptome. The translational profile of each gene displayed a cascade-like quality strikingly similar to the transcriptome (Figure 2B). Much like mRNA abundance, translation of phasic genes reaches a single maximum and a single minimum during the IDC. To determine the exact level of correlation between transcription and translation we directly compared mRNA and ribosome footprint density measurements (Figure 2C). In general, translation is tightly correlated with transcription for all phasic and non-phasic genes in rings ($r = 0.85$), early trophozoites ($r = 0.93$), late trophozoites ($r = 0.91$), schizonts ($r = 0.89$) and purified merozoites ($r = 0.86$). This indicates that when an mRNA is detected in one stage it is associated proportionally with ribosomes within the same stage. An example pair of genes is shown in Figure 3A. Here, mRNA abundance profiles of eukaryotic translation initiation factor eIF2 gamma subunit (PF14_0104) and the conserved protein PF14_0105, show that peak mRNA abundance for these two genes occurs at two different stages, early and late, respectively. Examination of ribosome occupancy of both genes reveals a ribosome density accumulation profile within the coding sequence that mirrors their respective mRNA profiles. As for the majority of genes, ribosome footprint density and mRNA abundance for these two genes are highly correlated (r = 0.98 and 0.93 for PF14_0104 and PF14_0105, respectively), indicating that mRNA translation occurs proportionally during the same stages at which these genes are transcribed (Figure 3B, supplementary file 1). Globally, 77% of genes expressed in at least 3 stages of the IDC display high Pearson correlation ($r \geq 0.7$) between mRNA abundance and translation (Figure 3-figure supplement 1). Thus, our genome-wide analysis of translation establishes that for the majority of genes expressed during the IDC, transcription and translation occur proportionally.

Ribosome profiling reveals instances of translational control of gene expression

Ribosome profiling allows the monitoring of translation rates through the simultaneous quantitative measure of mRNA abundance and ribosome density on mRNAs. The ratio of the footprint rpkM to the mRNA rpkM for any given gene represents its relative translational efficiency (TE) (Ingolia et al. 2009). To assess the dynamics of translational control and detect variations in control within and between developmental stages, we calculated the relative TE of all expressed genes in our dataset (Figure 4A). The shape and the range of TE distributions obtained for each stage sampled is comparable to those seen in other eukaryotes (Ingolia et al. 2009; Dunn et al. 2013). Absolute mean translational efficiencies in all stages ($\log_2$TE μ_{Rings} = -0.43, $\log_2$TE $\mu_{\text{E.trophs.}}$ = -0.56, $\log_2$TE $\mu_{\text{L.trophs}}$. = -0.31, $\log_2$TE $\mu_{\text{Schizonts}}$ = -0.16 and $\log_2$TE $\mu_{\text{Merozoites}}$ = -0.68) had a maximum fold difference of 1.47-fold observed between early trophozoites and schizonts. Translational efficiencies display a roughly 100-fold range in absolute values in each of the stages with the exception of the ring and merozoite stages, which exhibit more extreme values. In these stages the distribution of absolute TE values displays an approximately 4-fold larger spread than in early trophozoites, late trophozoites, or schizonts (Figure 4A, Figure 2-Source data 1). In rings the gene with the largest TE is the merozoite surface protein 9 (PFL1385c, $\log_2$TE = 4.1) and the gene with the lowest TE is the FIKK family serine/threonine protein kinase (PF14_0734, $\log_2$TE = -5.1). In merozoites the largest and lowest TE values correspond to the serine repeat antigen 5 (SERA5, PFB0340c, $\log_2$TE = 4.0) and the alpha adenylyl cyclase (PF14_0788a-c, $\log_2$TE = -4.7), respectively.

To determine the contribution of translational efficiency to the dynamic range of gene expression we examined the genes lying at the extremes of the TE distribution. For the purpose of this analysis, genes with a translational efficiency of two standard deviations above or below the mean in any of the stages were considered translationally up- or down-regulated, respectively. A total of 301 genes, 8.3% of the transcriptome, are translationally regulated by this metric, with 124 genes translationally down-regulated and 177 genes translationally up-regulated (Figure 4B, Figure 2-Source data 1). The timing of maximum mRNA expression does not influence TE for either of these two groups. Translational efficiencies remain high for the translationally up-regulated and low for the translationally down-regulated genes in all the stages at which they are expressed, regardless of the stage of peak mRNA abundance, suggesting that translational efficiency is largely, but not completely, programmed by the mRNA sequence itself, rather than global factors. For example, translational efficiency of the merozoite surface protein 6 (MSP6, PF10_0346) remains high ($\log_2$TE ≥ 2) across all stages irrespective of variations in its mRNA abundance. In contrast TE values for the eukaryotic initiation factor 2alpha kinase 1 (IK1, PF14_0423) are among the lowest measured despite high mRNA abundance across all stages (Figure 4C).

An examination of the 124 translationally down-regulated genes yielded some expected, and in some cases, unexpected findings. As would be expected, two pseudogenes, the ring-infected erythrocyte surface antigen 2 (RESA-2, PF11_0512) and reticulocyte binding protein homologue 3 (PfRh3, PFL2520w), represent a clear example of low translational efficiency. The PfRh3 pseudogene ribosome profile shows that translation of the 5’ end of this transcript occurs up until the encounter of several in-frame stop codons, causing the reduction in ribosome density from this point on (Figure 5A, Figure 5-figure supplement 1). This suggests that a truncated version of the PfRh3 protein is being produced in the W2 strain studied here. Evidence for peptides corresponding to the 5’ end of PfRh3 has been found in gametocytes and sporozoites (however not during the asexual stages) using mass spectrometry (Florens et al. 2002; Lasonder et al. 2002). We note that low levels of ribosomes can still be detected along the full length of this transcript in schizonts and merozoites. Whether these footprints derive from a low level of stop-codon read-through or accumulate via another unknown mechanism remains to be determined.

Ring-infected erythrocyte surface antigen 2, RESA2 (PF11_0512) was first described as a pseudogene based on the presence of an internal stop codon (Cappai et al. 1992). Since then, transcription of this gene has been demonstrated both *in vivo* (Vazeux et al. 1993) and *in vitro* (Bozdech et al. 2003). RESA-2 is transcribed but poorly translated in rings, early trophozoites and merozoites ($\log_2$TE -3.2, -2.7, -2.9, respectively). Accordingly, the ribosome profile of this gene in merozoites shows a general depletion of ribosomes along the CDS (Figure 5-figure supplement 2). In rings, ribosome density diminishes at the second exon. To validate the RESA2 gene model we used genomic DNA sequencing data derived from the *P. falciparum* W2 strain used in this study. We found that 69% (n = 151) of reads mapping to this locus support a single base deletion that creates a premature stop codon exactly at the site of ribosome footprint drop-off (Supplementary file 2). These data suggest that RESA2 is transcribed and actually translated into a shorter protein product of 461 amino acids. Whether or not the protein product is functional or undergoes post-translational degradation remains to be determined.

In addition to expected instances of translational regulation our data permits the discovery of previously uncharacterized translational regulation, especially at the extremes of the TE distributions. One of the most notable examples of translational silencing is the eIF2α kinase IK1 (PF14_0423) for which ribosome footprints accumulate at the 5' leader and 3' UTR but not on the CDS, resulting in an extremely low translational efficiency ($\log_2$TE = -3.6) despite relatively high transcript abundance across all stages (Figure 5B). The mechanism by which this gene is maintained in a translationally down-regulated state is unknown. Another example is the erythrocyte vesicle protein 1 (EVP1, PFD0495c) for which abundant transcript levels can be detected across all stages, with peak mRNA abundance occurring in rings and schizonts (Figure 5-figure supplement 2). Protein levels, however, have been shown to be undetectable (Tamez et al. 2008). Here, we find that translational efficiencies of this gene were low across all stages and lowest in rings and early trophozoites ($\log_2$TE -2.6 and -2.9, respectively) demonstrating that post-transcriptional regulation at the level of translation is, at least in part, responsible for its scarcity as a protein.

Thus, our ribosome profiling dataset highlights instances of translational control of genes that may not be detected by proteomic methods. Indeed a search for mass spectrometric data showed no evidence for ~70% of genes in this category (Aurrecoechea et al. 2009). Including the aforementioned examples, our dataset describes a total 124 translationally down-regulated genes (listed in Figure 2-Source data 1) for which translational efficiency values lie at the lower extremes of the distribution.

Protein products of translationally up-regulated genes are likely to be abundant and readily detected using mass-spectrometry. Previous proteomic studies show protein evidence in the blood stages for almost all (171 of 177) well-translated genes identified here (Pease et al. 2013; Aurrecoechea et al. 2009). Mass spectrometric evidence for the remaining 6 genes is either absent (PFL0245w, PFL2510w, PF11_0204) or has only been found in sporozoites (PFE1615c, MAL7P1.300, PF13_0069a). Despite the lack of proteomic data, our data indicate that these genes are both transcribed and translated during the blood stages of the parasite. Whether post-translational control points exist for these proteins is unknown.

Among the top ten most highly translated genes are proteins involved in merozoite egress and invasion MSP3, 6, 7 and 9 (merozoite surface proteins PF10_0345, PF10_0346, PF13_0197, and PFL1385c), serine repeat antigen 5 (SERA5, PFB0340c), and RAP1, 2, and 3 (PF14_0102, PFE0080c, and PFE0075c, respectively) (Figure 2-Source data 1). Interestingly, 73 (41%) of all translationally up-regulated genes can be assigned to the repertoire of canonical functions for merozoite egress and invasion described to date (Blackman 2008; Farrow et al. 2011; Yeoh et al. 2007; Hu et al. 2010). Strikingly, for all genes in this set maximum mRNA abundance is found during the late stages of the IDC (69 schizont and 4 merozoite stage mRNAs) yet for the majority (50, 70%) peak translational efficiency occurs in rings. Consistent with this, peptides for most of these merozoite function proteins (58 of 73) have been detected in rings (Oehring et al. 2012; Pease et al. 2013). This mode of translational regulation whereby late stage transcripts are highly translated in rings, was not exclusively limited to genes related to merozoite egress and invasion. We found evidence for an additional 14 genes with this profile, including, aquaglyceroporin (PF11_0338, $\log_2$TE = 3.8), tubulin beta chain (PF10_0084, $\log_2$TE = 1.8), and early transcribed membrane protein 2 (PFB0120w, $\log_2$TE = 2.5).

Taken together these data demonstrate that transcription and translation are tightly correlated for the majority of genes expressed during the asexual life cycle of *P. falciparum* with few exceptions. These apply to a small subset of translationally down- and up-regulated genes for which their translational efficiencies appear to be inherent properties of the mRNA, independent of changes in mRNA abundance. Genes in this category, especially those that exhibit high translational efficiencies, are enriched with functions associated with merozoite egress and invasion during the transition from late stages into rings.

Ribosome occupancy of 5' leaders is commonly found on genes expressed during the IDC

Ribosome profiling provides position specific information along each transcript allowing the detection of changes in ribosome distribution on the mRNA and their relationship to translational efficiency. To look for ribosome occupancy features beyond the CDS of transcripts, we first took advantage of the deep coverage and strand specificity of our RNA-seq data to identify 5' leaders and 3' UTRs of the *P. falciparum* transcriptome. We constructed a hidden Markov model (HMM) to automatically delineate the boundaries of both 5' leaders and 3' UTRs for known gene models (see materials and methods). Within the limits imposed by our data, we were able to describe 5' mRNA leaders and/or 3' UTRs for 3569 genes in at least one of the stages (Figure 6-figure supplement 1, Figure 2-Source data 1). 5' leaders are on average longer than 3' UTRs in each of the stages and median lengths across stages vary to a larger degree for 5' leaders (from 607 to 1040 nt) than for 3' UTRs (518 to 622 nt). The longest 5' mRNA leader was measured in late trophozoites (8229 nt) for the Ap2 transcription factor, PF11_0404, and the longest 3' UTR stretched 4773 nt for 60S ribosomal protein L7-3, PF14_0231, in rings. An example pair of genes with mapped 5' leaders and 3'UTRs is shown in Figure 6. Here, our HMM predicts a 636 nt and a 781 nt 5' leader and a 468 nt and 423 nt 3' UTR for the Myb2 transcription factor (PF10_0327) and the bromodomain protein (PF10_0328), respectively. These genes, encoded on opposite strands, share a 1536 nt intergenic sequence, however the span between the region delimited by their 5' leader sequence is only 120 nt and presumably harbors their respective promoters.

Next, using mRNA boundaries derived from our data, we analyzed ribosome distribution along each transcript during life cycle progression. More than 80% of the ribosome footprints in rings, early trophozoites, late trophozoites and schizonts, mapped to CDS regions of the genome, except in merozoites, where only 68% mapped to the CDS (Figure 7A). On average less than 1% of all reads obtained mapped to 3'UTRs in each stage, and most transcripts had no observed footprints past the stop codon. In contrast, footprints were far more common in 5' leaders (9.1%, 4.8%, 7.5%, and 4.8% in rings, early trophozoites, late trophozoites, and schizonts respectively) particularly in merozoites (23%). Footprint enrichment is specific to 5' leaders and not due to non-specific background since this would result in an increase of footprints mapping evenly along the length of the transcript, including the 3'UTR, and not just the 5' leader. Furthermore, these footprints most likely represent ribosomes because they derive from the 80S monosome fraction of the sucrose gradient, and their footprint read length distributions are equal to those of CDS mapping footprints, whereas they are significantly divergent from rRNA or tRNA read length distributions (Figure 7-figure supplement 1).

Upstream open reading frames are a major source for 5' leader ribosome density found from yeast to humans (Brar et al. 2012; Ingolia et al. 2009), and these have been shown to play a role in translational regulation of the downstream ORF in a few well-studied examples (Morris & Geballe 2000). In *P. falciparum* ribosomes have been suggested to accumulate on 5' leaders of genes displaying a delay in translation presumably due to long uORFs (Bunnik et al. 2013).

We defined 36,086 possible uORF regions in the 5' leaders of genes expressed during the *P. falciparum* IDC using a liberal definition that includes any stretch of at least 2 amino acids, starting with an AUG codon (Figure 7-Source data 1). Regardless of stage, half of the total ribosome footprint coverage in 5' leaders, in aggregate, or on a gene-by-gene basis did not overlap with these predicted uORFs (Figure 7B, Figure 7-figure supplement 2). We could find no significant correlation between the number of uORFs per gene, the uORF lengths, or the degree to which ribosome density was enriched in uORFs with translational efficiency (Figure 7-figure supplement 3). For example, erythrocyte binding antigen-175 (EBA-175, MAL7P1.176) is well translated in rings ($\log_2$TE = 1.4) and displays a large amount of 5' leader ribosome occupancy. Half (49%) of the reads map within the 9 predicted uORFs on the 5' leader of this gene, the other half maps outside these uORFs (Figure 7C). Using this liberal definition of a uORF, the data does not support an association between ribosome occupancy in these regions, nor does it support an association between the presence of these regions and translational efficiency.

Nonetheless, there exist at least two clear exceptions. First, we were able to validate translation of the reported uORF present in the 5' leader sequence of the var2csa mRNA which is expressed only in rings (Amulic et al. 2009). The majority of ribosome footprint density localizes to this uORF, and to a second one just upstream, while the var2csa coding sequence is largely devoid of footprints ($\log_2\mathrm{TE}_{\mathrm{Rings}}$= -4.2, Figure 7-figure supplement 4), consistent with its translational repression during growth in the absence of plancental tissue. Second, another striking example of uORF translation was found on PFE1550w (unknown function) for which the ratio of uORF to total 5' leader mapping reads is 0.9 (Figure 7-figure supplement 4). Indeed, ribosome footprint density is concentrated on one of the 6 uORFs predicted in the 5' leader of this gene, 56 amino acids long. This gene is also translationally down-regulated in all stages ($\log_2$TE = -2.7 on average). These two genes represent exceptional cases for which uORF translation negatively correlates with translation of the downstream ORF.

Aside from these two exceptions, for the vast majority of genes ribosome occupancy appears spread along 5' leaders, and not preferentially concentrated within possible uORFs. For this reason, we calculated 5' leader ribosome density (5'RD) for each gene expressed during the IDC, defined as upstream ribosome occupancy normalized for mRNA expression level and size of the leader sequence (5' leader ribosome footprint rpkM / 5' leader mRNA rpkM) (Figure 2-Source data 1). No positive correlation exists between the uORF and total 5' leader ribosome footprint coverage ratio, the number of uORFs per 5' leader, or their lengths, 5'RD, reinforcing the notion that uORFs are not a requisite for ribosome association to 5' leaders (Figure 7-figure supplement 5). In fact 5' ribosome density can be found on transcripts with 5' leaders completely devoid of AUGs and thus without uORFs by definition, such as the highly translated aquaglyceroporin ($\log_2$TE = 3.8 and $\log_2$5'RD = 2.9 in rings), and PFC0486c (unknown function, $\log_2$TE = 1.6 and $\log_2$5'RD = 1.1 in rings) (Figure 7-figure supplement 6).

Overall, rings and merozoite stage parasites were found to express transcripts with highest 5'RD (mean $\log_2$5'RD -0.03, and 0.11, respectively) relative to early trophozoites, late trophozoites and schizonts (mean $\log_2$5'RD -1.11, -0.26, -0.83), where the range of 5'RD values is also narrower (Figure 8A). Interestingly, among genes at the extremes of the 5'RD distributions (mean ± 1stdev) we also found many of our identified translationally up- and down-regulated transcripts (66 and 40%, respectively). On average, 5'RD was enriched on translationally up-regulated transcripts (mean $\log_2$5'RD = 0.83) and depleted for translationally down-regulated transcripts in all stages (mean $\log_2$5'RD = -1.11), suggesting the possibility that 5'RD is a byproduct of translational efficiency itself (Figure 8B).

In order to determine whether a direct relationship between 5'RD and translational efficiency of the downstream ORF exists, we compared these values for each gene. 5'RD positively correlates, albeit moderately, with translational efficiency in all stages, particularly in rings and merozoites (r = 0.51 and 0.49, respectively). We focused on the subset of genes with highest and lowest 5'RD values (mean ± 2 stdev) and found that only a fraction of the translationally up- and down-regulated genes overlap with this category of extreme 5'RDs in each stage (Figure 8C). The largest overlap occurred in rings where the highest 5'RD values were found in 43% (31 genes) of the translationally up-regulated genes, including MSP6, AQP and SERA5. These results indicate that while in general a correspondence between 5'RD and translational efficiency exists, one is not necessarily predictive of the other and exceptions apply. This is the case, for example, of the translationally down-regulated transcript of the pseudogene PfRh3, which in rings has the second highest 5'RD value ($\log_2$5'RD = 5.1).

In summary our data establishes ribosome accumulation on 5' leaders as a common feature of transcripts expressed during the IDC. Ribosome density is not restricted to predicted uORFs present within these regions and, with few exceptions, the uORF number, length, or coverage level, is not a requirement for 5' ribosome density and has no measurable effects on the translation of the downstream ORF. Even though 5'RD is more commonly found on 5' leaders of highly translated transcripts, this is not a universal trend since only a moderate correlation exists between 5'RD and the translational efficiency of the downstream ORF.

3'UTR ribosome occupancy is rare

While our data showed 3'UTRs to be relatively depleted of ribosomes, we searched for rare cases of high 3'UTR ribosome density, possibly arising from stop codon read-through, alternative stop codon usage, or re-initiation of downstream ORFs (Guydosh & Green 2014; Dunn et al. 2013). We systematically searched for transcripts for which coverage, in a sliding window of 30 nt, was greater in the 3'UTR than the CDS, and found 19 genes meeting this criterion. These genes could be qualitatively divided into two categories: 14 with putative stop codon read-through and/or alternate stop codon usage, and 5 genes for which the origin of the 3'UTR density is unclear (listed in Figure 9-Source data 1). An example of stop codon read-through is the conserved plasmodium protein (PF13_0160), shown in Figure 9A. Ribosomes not only extend beyond the annotated stop codon of this transcript but also skip subsequent in-frame stop codons present on the predicted 644 nt 3'UTR. Interestingly, ribosome footprints accumulate in a single large peak approximately 130 nt downstream of the annotated stop codon. On the 1290 nt 3'UTR of the sodium-dependent phosphate transporter (MAL13P1.206), two large peaks of ribosome footprint density, one approximately 560 nt and the other 860 nt from the stop codon, can be observed (Figure 9B). The origin of these footprints is unclear and it is possible that these are the product of nuclease protection by RNA-binding proteins that co-sediment with the 80S monosome. To confirm that 3'UTR mapping reads are derived from ribosome footprints we compared their cumulative read length distributions against a typical CDS footprint read length distribution (Figure 9-figure supplement 1). For the 16 of the 19 genes we observed no significant difference in footprint size distributions localized to the CDS compared to the 3'UTR. For the remaining three genes, the sodium-dependent phosphate transporter (MAL13P1.206), the acyl-Coa synthetase (PFD0085c), and the conserved plasmodium protein (PF13_0160), 3'UTR footprint size distributions were divergent from that of the CDS, implying that footprints found on these genes' 3'UTRs may be produced by nuclease protection of these regions by factors other than ribosomes and that co-sediment with 80S ribosomes.

Antisense detection

Antisense transcription plays an important role in gene regulation from bacteria to humans and while its role is increasingly studied in these organisms (Faghihi & Wahlestedt 2009), less is known about its relevance in *P. falciparum*. Previous serial analysis of gene expression (SAGE) (Patankar et al. 2001), nuclear run-on experiments (Militello et al. 2005) and more recently antisense splicing events detected by RNA-seq (Sorber et al. 2011; López-Barragán et al. 2011), suggest that antisense RNAs are synthesized by RNA pol II and may constitute up to ~12% of the erythrocytic-stage steady-state RNA (Gunasekera et al. 2004), yet their presence and biological role, if any, remains unclear. A more recent study found no correlation between natural antisense transcript levels and protein abundance (Siegel et al. 2014).

The 30 nt fragmentation and RNA-ligase based library preparation method employed here affords exquisite strand specificity by minimizing artifacts associated with random priming during reverse transcription. As evidence of this specificity, the highest expressed gene in our data set, histone h2a (PFF0860c), yielded a total of 765,510 reads on the sense strand, and only 2 reads on the antisense strand, corresponding to a sense:antisense ratio greater than 10^5. Furthermore, our HMM mapping of 5' leaders and 3' UTRs facilitates the differentiation between independently transcribed antisense RNA and transcripts that occur by virtue of being part of an adjacent gene. We took advantage of the nature of our dataset to identify antisense transcripts and looked for effects on sense mRNA translation.

For this analysis only, we relaxed our stringent coverage threshold from ≥32 rM to ≥16 rM for inclusion of antisense transcripts. We based our threshold on the presence of an antisense transcript to the sexual stage specific gene pfs16 (PFD0310w) confirmed by strand specific RT-PCR (Figure 10-figure supplement 1). This antisense is predicted by the HMM to be ~4 kb, extending over the complete coding sequence and beyond, and with a coverage level of 23 rM over the sense CDS. Using the 16 rM threshold we detected 84 antisense transcripts to several known ORFs (listed in Figure 10-Source data 1), including the nucleoside transporter pfNT4 (PFA0160c) depicted in Figure 10A. The merozoite stage contained the highest number of antisense transcripts (46) and the fewest (13) were found in early trophozoites. Manual inspection revealed that in 63% of these instances, the putative antisense transcript actually emanates from the 5' leader or 3' UTR of a neighboring gene (not defined by the HMM). Antisense reads for the para-hydroxybenzoate polyprenyltransferase (PFF0370w) for example, are actually derived from the 3'UTR of the neighboring conserved protein PFF0375c (Figure 10B).

We next interrogated the impact of this set of antisense transcripts. Overall, antisense transcripts showed no effect on mRNA abundance and translational efficiencies of the cognate sense transcript. These observations parallel those described for antisense transcripts in yeast (Brar et al. 2012). Thus, at first approximation antisense transcripts do not appear to play a role in translational regulation. However, these observations could be confounded due to the small number of genes in this set and we cannot exclude the possibility of sense/anti-sense heterogeneity at the single cell level, obscured here at the population level.

Discussion

Herein we present, for the first time, a comprehensive view of the coupled transcriptional and translational dynamics of the *P. falciparum* IDC by determination of transcript abundance and architecture together with ribosomal density and positioning. The quality of our data relies on several critical features: 1) high temporal specificity and reproducibility of fully independent biological replicas of five strictly staged cultures 2) purified merozoites to allow discrete measurements in this stage without confounding contributions from schizonts or rings 3) monosome isolation from sucrose gradients to specifically enrich for ribosome-derived footprints and avoid potential complications that can arise with methods like sucrose cushions which are prone to mRNA contamination 4) sufficient sequencing depth of biological replicates to set a statistical threshold for minimum read coverage and to demonstrate reproducibility 5) stringent strand specificity to facilitate an HMM for the description of transcript boundaries and the detection of antisense transcription.

Previous studies of the transcript abundance in the malaria blood stages revealed a periodic cascade of gene expression, whereby the majority of expressed genes exhibit one peak of expression per cell cycle (Bozdech et al. 2003). The global profile of transcriptional expression was subsequently found to be highly stereotypical across strains, and appeared to lack dynamic responses to perturbation (Llinás et al. 2006; Ganesan et al. 2008). It has been suggested that translational control of protein expression could compensate for the lack of transcriptional dynamics. Proteomic studies described delays in peak mRNA and corresponding protein abundance implicating translational or post-translational mechanisms in the modulation of gene expression (Foth et al. 2011; Le Roch et al. 2004).

Our ribosome profiling results reveal a tight coupling of transcription and translation for the majority of expressed genes, indicating that most protein products are translated with highly similar timing and in proportion to their corresponding mRNA transcripts. Synthesized proteins are likely to exert their functions immediately upon translation but post-translational regulation, not captured by our data, could still be at play. Direct correlations of translational efficiencies measured in this study along with proteomic datasets are hampered by the reduced sensitivity of the latter, and differences in temporal resolution and staging of the parasites between datasets. However, the available proteomic evidence is largely consistent with the results presented here, particularly for highly translated proteins. The simultaneous capture of mRNA abundance and translation is expected to be a more accurate proxy for protein levels than measurements of mRNA abundance alone (Ingolia et al. 2009) and provides a critical resource for the identification of instances of post-translational regulation of gene expression. However, we note that this dataset only provides a direct measure of relative changes in translational efficiencies rather than changes in bulk transcription and translation.

While no up- or down- regulation of global translation efficiencies were observed in any particular stage, more extreme translational efficiencies were measured in subsets of genes expressed in rings and merozoites. We find 177 translationally up-regulated genes with functions predominantly related to merozoite egress and invasion, with peak mRNA in schizonts and peak translational efficiency in rings. It is likely that the genes with unknown functions, regulated in an analogous way during the merozoite to ring transition, are also associated with this process. Our data supports a model whereby the transcripts of proteins necessary for merozoite structure and function are made in the previous stage in large abundance, are translationally up-regulated during the invasion process, and remain highly translated well into the ring stage despite rapid mRNA decay during this stage (Shock et al. 2007). Whether the accumulation of 5' leader ribosome density is a mechanism that assists in this process or is it merely a byproduct of more efficient ribosomal initiation on these templates remains to be tested. With the emergence of genome editing tools such as CRISPR/Cas9 (Ghorbal et al. 2014) it may be possible to create versions of genes with altered cis-acting sequences to test for modulation of 5' ribosome density and its effect on translational efficiency.

The global nature of ribosome accumulation within the 5' leader sequences of many transcripts during the IDC and the lack of an association between 5'RD and the number or length of uORFs suggests that ribosomes accumulate on 5' leaders through means other than a uORF model. For comparison, in yeast under starvation conditions the fraction of ribosome footprints derived from 5' leaders is increased by six-fold and in some cases no single uORF can account for the entire distribution of ribosomes on the 5' leader of a gene (Ingolia et al. 2009), much like *P. falciparum*. What mechanism could account for global ribosome accumulation in the 5' leader? The presence of apparent 80S ribosomes within the 5' leader sequence, regardless of whether they cover uORFs or not, suggests an engagement mode in which the fidelity of start codon recognition is altered or suspended. Current models propose that the 43S pre-initiation complex loads onto the mRNA with the assistance of other initiation factors near the 5' cap, and proceeds to scan down the length of the mRNA until it encounters an AUG codon. This is followed by assembly of the 48S preinitiation complex and then finally the 80S complex (for review, see Hinnebusch 2011). The AUG that is ultimately chosen is not always the first one encountered, and its sequence context is important for recognition. The factors eIF1, eIF1A, and eIF5 have been implicated in recognition of the "correct" AUG (Aitken & Lorsch 2012). In the case of *P. falciparum*, differential regulation or modification of these factors could plausibly result in altered start codon selection and 80S assembly. Whether prematurely initiated complexes are able to scan without synthesizing a peptide or are required to assemble and reassemble until encountering the right start codon remains an open question. Large 5' ribosome accumulation on translationally up-regulated genes in the ring stage suggests that premature initiation on these transcripts is not detrimental. The development of an *in vitro* translation system that recapitulates upstream 80S assembly on *P. falciparum* 5' leaders will allow direct testing of premature initiation and its effect on translational efficiency in this parasite.

Our ribosome profiling data adds an important component to the rich compendium of genome-wide data, including transcript abundance (Bozdech et al. 2003), mRNA decay (Shock et al. 2007), splicing (Sorber et al. 2011), and proteomics for this parasite (Foth et al. 2011; Le Roch et al. 2004). Features such as 5' leaders, 3'UTRs, introns, and antisense transcripts are clearly visible and often well delineated. While experimental validation of transcriptional start sites, terminators, and promoters is required, spanning regions between transcripts, such as the one shown in Figure 6, can be used for the search and identification of such functional sites in a reduced sequence space. The data is available at NCBI GEO (accession #GSE58402) to facilitate future queries and normalized read coverage plots for all 5 timepoints are available packed as a single Mochiview file (ADD REFERENCE). Together our results describe a simplified regulatory architecture of gene translation, albeit one that includes peculiar and potentially unique mechanisms specialized for its highly structured and coordinated lifecycle within erythrocytes. Further biochemical dissection of translational initiation mechanisms and determinants of translational efficiency unique to *Plasmodium* may reveal weaknesses that could be exploited for possible therapeutic intervention.

Materials and Methods

Cell culture

W2 strain cultures were maintained in Hyperflasks (Corning) in 500 ml RPMIc (RPMI 1640 media supplemented with 0.25% Albumax II (GIBCO), 2 g/l sodium bicarbonate, 0.1 mM hypoxanthine, 25 mM HEPES (pH 7.4), and 50 μg/l gentamycin), at 37°C, 5% O_2, and 5% CO_2, to maximum 10-15% parasitemia at 5% hematocrit (HC) and frequent media changes (at least every 6-8 hours). Cells were synchronized by two consecutive sorbitol treatments for three generations, for a total of six treatments. Maximum invasion, point at which half of the culture is either rings or schizonts, was defined as hour zero and independent time points containing ~10^{10} parasites were harvested 11, 21, 31 and 45 hours later.

Polysome isolation and library generation

Cultures were incubated for 5 min in 500 ml 37ºC RPMIc, 100μg/ml cycloheximide (Acros Organics) and harvested by centrifugation for 8 min at 3.65 xg at room temperature. An aliquot was removed and flash frozen in liquid nitrogen for total total RNA purification, followed by poly(A)-purification and chemical fragmentation with Zn^{2+} to ~30 nt for consistency in mRNA-Seq library preparation. The remaining culture was treated with ice-cold 0.1% saponin in 1X PBS, 100ug/ml cycloheximide, for RBC lysis. Parasites were resuspended in ice-cold parasite lysis buffer (15 mM KOAc, 15 mM MgOAc, 10 mM Tris HCl pH 7.4, 0.5 mM DTT, 0.5% Triton X-100, 100 ug/ml cycloheximide) and dripped into a conical tube filled with, and immersed in, liquid nitrogen. Frozen cells transferred placed in liquid nitrogen pre-chilled chambers and pulverized for 3 min at 15 Hz, on a Retsch MM301 mixer mill. Pulverized cells were thawed on ice and cell debris was removed by centrifugation at 4ºC, 16000 xg for 10 min. The supernatant was treated with 2.88 U/ug Micrococcal nuclease for 30 min at room temperature and immediately loaded onto sucrose gradients for ultracentrifugation at 35000 rpm for 3 h at 4°C in a L8-60M Beckman centrifuge. Monosome fractions only, were collected to generate ribosome footprint libraries for deep sequencing using the HiSeq 2000 (Illumina), as described (Ingolia et al. 2009).

Merozoite Purification

Late stage schizonts (40-44 hpi) were magnetically purified using MACS LD columns (Miltenyi Biotec, San Diego, CA) and resuspended in RPMIc without blood addition. After reaching maximum invasion (1:1 schizont to ring ratio) cultures were harvested by centrifugation at 1500 rpm at room temperature for 5 min. Pelleted cultures were resuspended into fresh RPMIc and placed at 37°C. Merozoites in the supernatant were treated with 100ug/ml cycloheximide for 5 min at room temperature, harvested at 4000rpm at 4°C for 5 min and resuspended in RPMIc and passaged again through a MACS LD column. Parasite lysis buffer was added to the merozoite-enriched flow-through and flash frozen in liquid nitrogen. This procedure was repeated three times every 45 min using the original culture. The same procedure described above was used for RNA extraction and library preparation.

SNP-corrected W2 genome

W2 strain genomic DNA was isolated from >90% synchronized ring stage cultures. Paired end libraries were constructed using the Nextera DNA Sample Prep Kit (Epicenter Biotechnologies, Madison, WI) according to the manufacturer's instructions reducing PCR cycles from 9 to 6 and using 80% A/T dNTPs. Libraries were sequenced using the HiSeq 2000 (Illumina, San Diego, CA). Reads were aligned to the *P. falciparum* PlasmoDB 3D7 version 7.1 genome using Bowtie 0.12.1 (Langmead et al. 2009) with parameters –v1 –m 1 (one mismatch allowed, unique mapping). A SNP was called when 5 or more W2 reads supported, with over 90% agreement, a different base than the one found in the *P. falciparum* 3D7 7.1 genome. 19401 SNPs (0.08% of total bases) were detected and used to produce the SNP-corrected W2 genome based on the 3D7 genome. Fastq files are available for download at NCBI SRA, accession #SRP042946.

Software Pipeline, mappability and rpkM calculation

Quality-filtered ribosome footprints and mRNA sequencing reads were trimmed to remove library adapter sequence, filtered for *P. falciparum* rRNA using blast, and aligned uniquely to the W2 SNP-corrected genome using Bowtie 0.12.1 (Langmead et al. 2009) allowing no mismatches. The percentage mappability was calculated using an *in silico* library of the *P. falciparum* W2 SNP-corrected genome created using a single nucleotide sliding window of 30nt. The *in silico* library was uniquely aligned to the genome allowing no mismatches. The mappability score is given by the number of 30 nt sequences covering each nt position in the genome, such that any position has a score that ranges from 0 to 30. Both mRNA and ribosome footprint rpkMs were calculated as in (Mortazavi et al. 2008), excluding the first 50 bases of each gene to eliminate bias introduced by the observed ribosome accumulation peak near the start codon. Genes with fewer than 80% mappable bases (248 genes) or any overlapping non-CDS feature on the same strand (77 genes) were excluded from this calculation. Data is available for download at NCBI GEO, accession #GSE58402. MochiView genome browser data tracks are available in supplementary file 1 (Homann et al., 2010).

Extended phaseogram

The genes of the RNA-seq transcriptome obtained in this work were listed in the same phaseogram order as the previously published microarray transcriptome (Bozdech et al. 2003).

The criteria for inclusion of a gene into the phaseogram was mRNA ≥ 32 rM, >2 peak to trough ratios, and Pearson correlation coefficient >0.8 with the expression profiles of the two neighboring genes.

Hidden Markov Model to describe transcript boundaries

The HMM was built using RNA-Seq data obtained in this study and two states: transcript (*t*) or intergenic (*i*) with three possible emissions: 1) read present, 2) read not present but position is unmappable, 3) read not present but the position is mappable. Both state and emission probabilites were calculated using a ~30kb training set of manually identified transcript and non-transcript regions. The initial probabilities were set to 0.5. Transition probabilities were estimated from the median length of intergenic regions of (1252 nt) and median lengths of CDS regions (2545 nt), where the $P_{t\text{->}i}$ = (1/2545), $P_{t\text{->}t}$ = (2544/2545), $P_{i\text{->}t}$ = (1251/1252), and $P_{i\text{->}i}$ = (1/1252). We applied the Viterbi algorithm to predict the optimal path of transcript tracks per time point with a 10 nt window resolution. HMM-defined 5' leader and 3' UTR coordinates are available for download at NCBI GEO, accession #GSE58402.

Strand-specific RT-PCR

Total RNA from late stage parasites was isolated and reverse transcribed using SuperScript III (Invitrogen, Carlsbad, CA) according to manufacturer's instructions, using either an antisense-specific primer to Pfs16 (PFD0310w) or a random nonamer. cDNA was amplified using the Pfs16 antisense-specific primer as a forward primer in combination with one of five reverse primers (Supplementary file 3). 18S rRNA primers were used in the control reactions with the random nonamer derived cDNA.

Acknowledgments

The authors would like to thank Dr. Graham Ruby for technical expertise in devising the HMM for UTR description. We would also like to thank members of the Weissman Lab Dr. Eugene Oh, and Dr. Gloria Brar for helpful experimental advice and suggestions for the analysis of data and Dr. Jonathan Weissman for the use of a density gradient fractionator and for helpful discussions during article preparation.

References

Aitken, C.E. & Lorsch, J.R., 2012. A mechanistic overview of translation initiation in eukaryotes. *Nature structural & molecular biology*, 19(6), pp.568–576. doi:10.1038/nsmb.2303.

Amulic, B. et al., 2009. An upstream open reading frame controls translation of var2csa, a gene implicated in placental malaria. *PLoS pathogens*, 5(1), p.e1000256. doi:10.1371/journal.ppat.1000256.

Aurrecoechea, C. et al., 2009. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic acids research*, 37 (Database issue), pp.D539–543. doi:10.1093/nar/gkn814.

Blackman, M.J., 2008. Malarial proteases and host cell egress: an "emerging" cascade. *Cellular Microbiology*, 10(10), pp.1925–1934. doi:10.1111/j.1462-5822.2008.01176.x.

Bozdech, Z. et al., 2003. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biology*, 1(1), p.E5. doi:10.1371/journal.pbio.0000005.

Brar, G.A. et al., 2012. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science (New York, N.Y.)*, 335(6068), pp.552–557. doi:10.1126/science.1215110.

Bunnik, E.M. et al., 2013. Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum. *Genome biology*, 14(11), p.R128. doi:10.1186/gb-2013-14-11-r128.

Cappai, R. et al., 1992. Cloning and analysis of the RESA-2 gene: a DNA homologue of the ring-infected erythrocyte surface antigen gene of Plasmodium falciparum. *Molecular and Biochemical Parasitology*, 54(2), pp.213–221. doi:10.1016/0166-6851(92)90113-X.

Dunn, J.G. et al., 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife*, 2, p.e01179. doi:10.7554/eLife.01179.

Faghihi, M.A. & Wahlestedt, C., 2009. Regulatory roles of natural antisense transcripts. *Nature Reviews. Molecular Cell Biology*, 10(9), pp.637–643. doi:10.1038/nrm2738.

Farrow, R.E. et al., 2011. The mechanism of erythrocyte invasion by the malarial parasite, Plasmodium falciparum. *Seminars in cell & developmental biology*, 22(9), pp.953–960. doi:10.1016/j.semcdb.2011.09.022.

Florens, L. et al., 2002. A proteomic view of the Plasmodium falciparum life cycle. *Nature*, 419(6906), pp.520–526. doi:10.1038/nature01107.

Foth, B.J. et al., 2011. Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite Plasmodium falciparum. *Molecular & cellular proteomics: MCP*, 10(8), p.M110.006411. doi:10.1074/mcp.M110.006411.

Ganesan, K. et al., 2008. A genetically hard-wired metabolic transcriptome in Plasmodium falciparum fails to mount protective responses to lethal antifolates. *PLoS pathogens*, 4(11), p.e1000214. doi:10.1371/journal.ppat.1000214.

Gardner, M.J. et al., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906), pp.498–511. doi:10.1038/nature01097.

Ghorbal, M. et al., 2014. Genome editing in the human malaria parasite Plasmodium falciparum using the CRISPR-Cas9 system. *Nature biotechnology*. doi:10.1038/nbt.2925.

Gunasekera, A.M. et al., 2004. Widespread distribution of antisense transcripts in the Plasmodium falciparum genome. *Molecular and biochemical parasitology*, 136(1), pp.35–42. doi:10.1016/j.molbiopara.2004.02.007.

Guydosh, N.R. & Green, R., 2014. Dom34 rescues ribosomes in 3' untranslated regions. *Cell*, 156(5), pp.950–962. doi:10.1016/j.cell.2014.02.006.

Hinnebusch, A.G., 2011. Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiology and molecular biology reviews: MMBR*, 75(3), pp.434–467, first page of table of contents. doi:10.1128/MMBR.00008-11.

Homann, O.R. et al., 2010. MochiView: versatile software for genome browsing and DNA motif analysis. Biomed Central Biology, 8(49), doi: 10.1186/1741-7007-8-49

Hu, G. et al., 2010. Transcriptional profiling of growth perturbations of the human malaria parasite Plasmodium falciparum. *Nature Biotechnology*, 28(1), pp.91–98. doi:10.1038/nbt.1597.

Hughes, K.R. et al., 2010. From cradle to grave: RNA biology in malaria parasites. *Wiley interdisciplinary reviews. RNA*, 1(2), pp.287–303. doi:10.1002/wrna.30.

Ingolia, N.T. et al., 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, 324(5924), pp.218–223. doi:10.1126/science.1168978.

Ingolia, N.T., 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics*, 15(3), pp.205–213. doi:10.1038/nrg3645.

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), p.R25. doi:10.1186/gb-2009-10-3-r25.

Lasonder, E. et al., 2002. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature*, 419(6906), pp.537–542. doi:10.1038/nature01111.

Le Roch, K.G. et al., 2004. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome research*, 14(11), pp.2308–2318.doi:10.1101/gr.2523904.

Llinás, M. et al., 2006. Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Research*, 34(4), pp.1166–1173. doi:10.1093/nar/gkj517.

López-Barragán, M.J. et al., 2011. Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. *BMC genomics*, 12, p.587. doi:10.1186/1471-2164-12-587.

Mair, G.R. et al., 2006. Regulation of sexual development of Plasmodium by translational repression. *Science (New York, N.Y.)*, 313(5787), pp.667–669. doi:10.1126/science.1125129.

Mair, G.R. et al., 2010. Universal features of post-transcriptional gene regulation are critical for Plasmodium zygote development. *PLoS pathogens*, 6(2), p.e1000767. doi:10.1126/science.1125129.

Militello, K.T. et al., 2005. RNA polymerase II synthesizes antisense RNA in Plasmodium falciparum. *RNA (New York, N.Y.)*, 11(4), pp.365–370. doi:10.1261/rna.7940705.

Morris, D.R. & Geballe, A.P., 2000. Upstream Open Reading Frames as Regulators of mRNA Translation. *Molecular and Cellular Biology*, 20(23), pp.8635–8642. doi:10.1128/MCB.20.23.8635-8642.2000.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), pp.621–628. doi:10.1038/nmeth.1226.

Nirmalan, N., Sims, P.F.G. & Hyde, J.E., 2004. Quantitative proteomics of the human malaria parasite Plasmodium falciparum and its application to studies of development and inhibition. *Molecular microbiology*, 52(4), pp.1187–1199. doi:10.1016/j.molbiopara.2004.02.013.

Nirmalan, N., Sims, P.F.. & Hyde, J.E., 2004. Translational up-regulation of antifolate drug targets in the human malaria parasite Plasmodium falciparum upon challenge with inhibitors. *Molecular and Biochemical Parasitology*, 136(1), pp.63–70.

Oehring, S.C. et al., 2012. Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum. *Genome biology*, 13(11), p.R108. doi:10.1186/gb-2012-13-11-r108.

Patankar, S. et al., 2001. Serial Analysis of Gene Expression in Plasmodium falciparum Reveals the Global Expression Profile of Erythrocytic Stages and the Presence of Anti-Sense Transcripts in the Malarial Parasite. *Molecular Biology of the Cell*, 12(10), pp.3114–3125. doi: 10.1091/mbc.12.10.3114

Pease, B.N. et al., 2013. Global analysis of protein expression and phosphorylation of three stages of Plasmodium falciparum intraerythrocytic development. *Journal of proteome research*, 12(9), pp.4028–4045. doi:10.1021/pr400394g.

Shock, J.L., Fischer, K.F. & DeRisi, J.L., 2007. Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome biology*, 8(7), p.R134. doi:10.1186/gb-2007-8-7-r134.

Siegel, T.N. et al., 2014. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. *BMC genomics*, 15, p.150. doi: 10.1186/1471-2164-15-150.

Sorber, K., Dimon, M.T. & DeRisi, J.L., 2011. RNA-Seq Analysis of Splicing in Plasmodium Falciparum Uncovers New Splice Junctions, Alternative Splicing and Splicing of Antisense Transcripts. *Nucleic Acids Research*, 39(9), pp.3820–3835. doi:10.1093/nar/gkq1223.

Tamez, P.A. et al., 2008. An Erythrocyte Vesicle Protein Exported by the Malaria Parasite Promotes Tubovesicular Lipid Import from the Host Cell Surface. *PLoS Pathogens*, 4(8). doi:10.1371/journal.ppat.1000118

Vazeux, G., Scanf, C.L. & Fandeur, T., 1993. The RESA-2 gene of Plasmodium falciparum is transcribed in several independent isolates. *Infection and Immunity*, 61(10), pp.4469–4472.

Yeoh, S. et al., 2007. Subcellular Discharge of a Serine Protease Mediates Release of Invasive Malaria Parasites from Host Erythrocytes. *Cell*, 131(6), pp.1072–1083. doi:10.1016/j.cell.2007.10.049

Zhang, K. & Rathod, P.K., 2002. Divergent Regulation of Dihydrofolate Reductase Between Malaria Parasite and Human Host. *Science*, 296(5567), pp.545–547. doi:10.1126/science.1068274

Zhang, M. et al., 2012. PK4, a Eukaryotic Initiation Factor 2α(eIF2α) Kinase, Is Essential for the Development of the Erythrocytic Cycle of Plasmodium. *Proceedings of the National Academy of Sciences*, 109(10), pp.3956–3961. doi:10.1073/pnas.1121567109.

Zhang, M. et al., 2010. The Plasmodium Eukaryotic Initiation Factor-2α Kinase IK2 Controls the Latency of Sporozoites in the Mosquito Salivary Glands. *The Journal of Experimental Medicine*, 207(7), pp.1465–1474. doi:10.1084/jem.20091975.

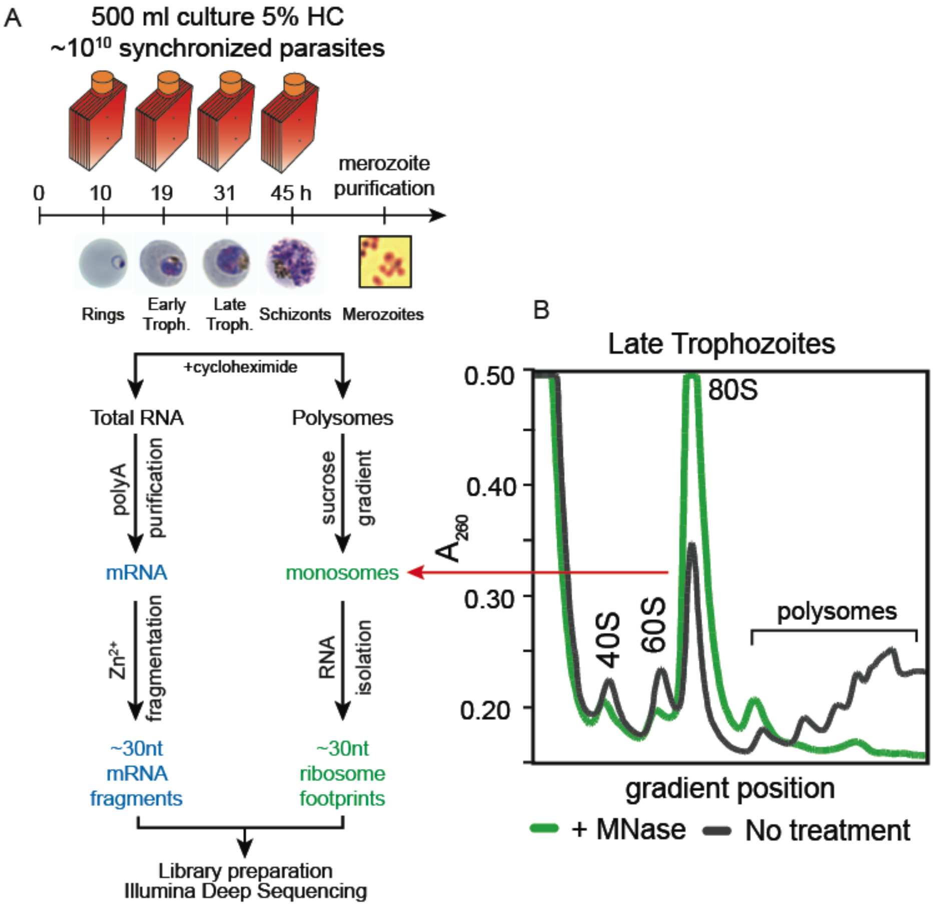


Figure 1. Ribosome profiling of the *P. falciparum* asexual blood stages, experimental outline. (**A**) Synchronized parasite cultures were maintained in hyperflasks at 5% hematocrit and maximum 15% parasitemia. Cycloheximide-treated cultures containing ~1010 parasites were harvested at ring, early trophozoite, late trophozoite and schizont stages (10, 21, 31 and 45 hpi, respectively) for total RNA or polysome isolation. Merozoites were purified through magnetic isolation of late stage schizonts (see materials and methods). Nuclease treated polysomes were fractionated on a sucrose gradient. Ribosome footprints (~30 nt) derived from the monosome peak (dashed red line) or chemically fragmented polyA purified mRNA (~30 nt) were used to build sequencing libraries. mRNA and ribosome footprint samples were processed in parallel to create deep sequencing libraries compatible with the Illumina platform. (**B**) Sucrose gradient A_{260} absorbance profile of polysome extracts derived from late trophozoites treated with micrococcal nuclease (green, +MNase) or untreated controls (gray, No treatment). Red arrow indicates the 80S monosome peak collected for ribosome footprint library preparation.

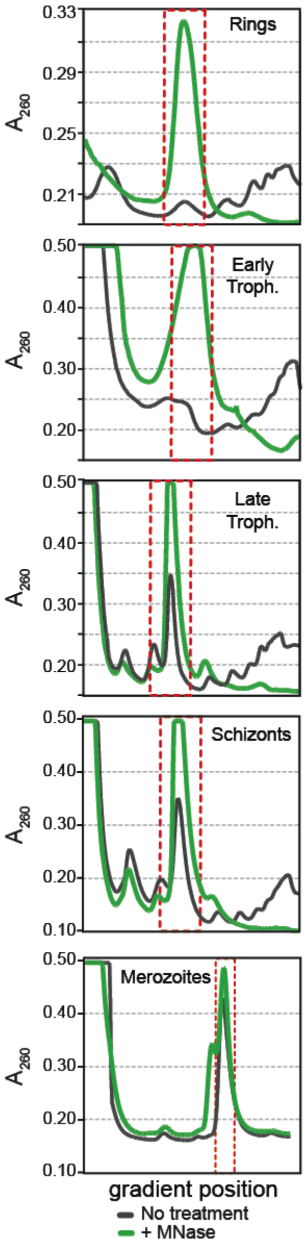


Figure 1-figure supplement 1. Polysome profiles of the *P. falciparum* asexual blood stages. Sucrose gradient A_{260} absorbance profiles of polysome extracts treated with micrococcal nuclease (green, +MNase) and untreated controls (gray, No treatment). Red dotted line indicates monosome peak harvested for ribosome footprint library generation.

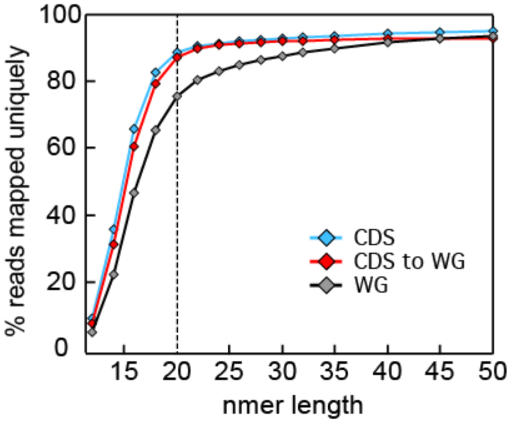


Figure 1-figure supplement 2. Read size influence on mappability. Single nucleotide sliding windows ranging from 10 to 50 nt were used to generate in silico libraries of the *P. falciparum* W2 SNP corrected genome. These were uniquely aligned, allowing no mismatches, to either the whole genome (gray, WG) or the coding sequences (blue, CDS) using Bowtie (Mortazavi et al. 2008) and the percentage of aligned reads were calculated for each window size. The analysis was repeated using sliding windows generated from the coding genome only (red, CDS to WG) for a more representative mappability estimate of an RNA-seq dataset. Read sizes of ≥20 nt asymptotically approach maximum mappability percentages.

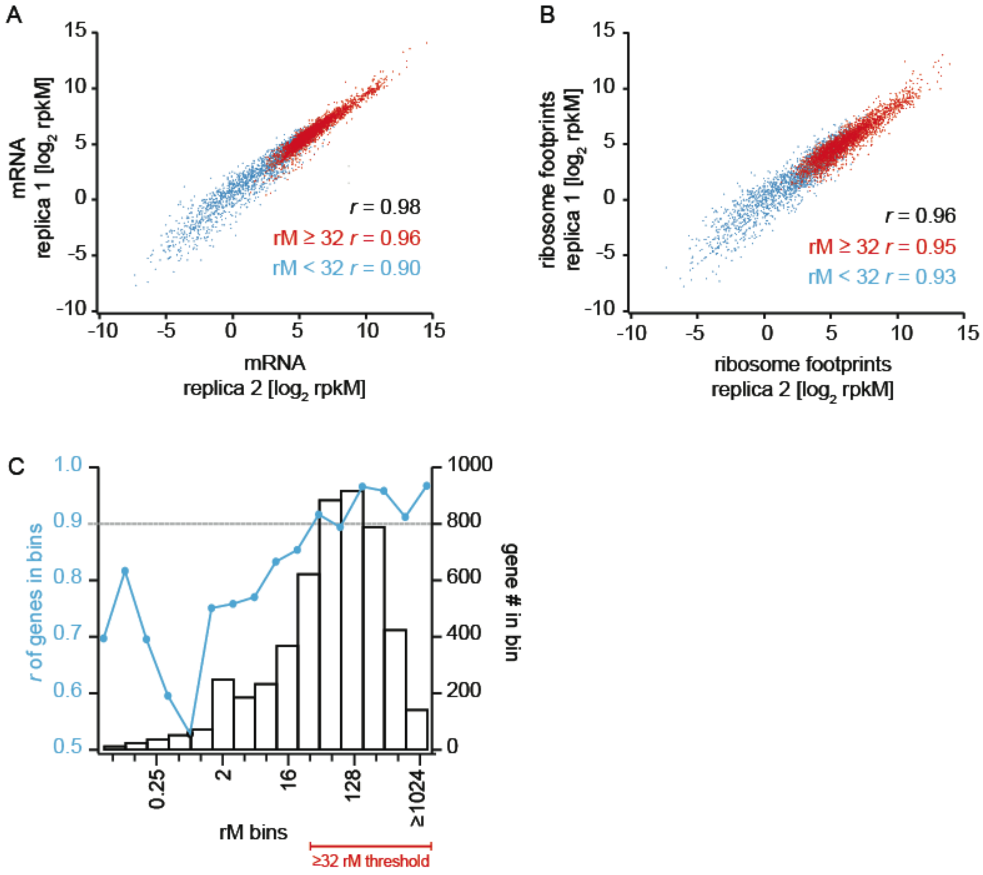


Figure 1-figure supplement 3. Reproducibility and coverage threshold determination using two fully independent biological replicates. mRNA abundance measurements (**A**) and ribosome footprint densities (**B**) in terms of rpkM in two fully independent biological replicates of the late trophozoite timepoint. Genes with at least ≥32 total mRNA reads counted (rM) are highly reproducible (r ≥ 0.9) across replicas (A and B red dots, and Figure D) whereas low read counts have a negative effect on rpkM reproducibility (A and B blue dots, and C). (**C**) Genes were binned based on their rM in replica 1. In each bin Pearson correlations of rpkM values of replica 1 and replica 2 were calculated. At 32 rM *r* values were consistently above 0.9 indicating that rpkMs calculated for genes with ≥32 rM are highly reproducible across replicates, and this is independent of the number of genes in the bin. *r* = Pearson correlation coefficient.

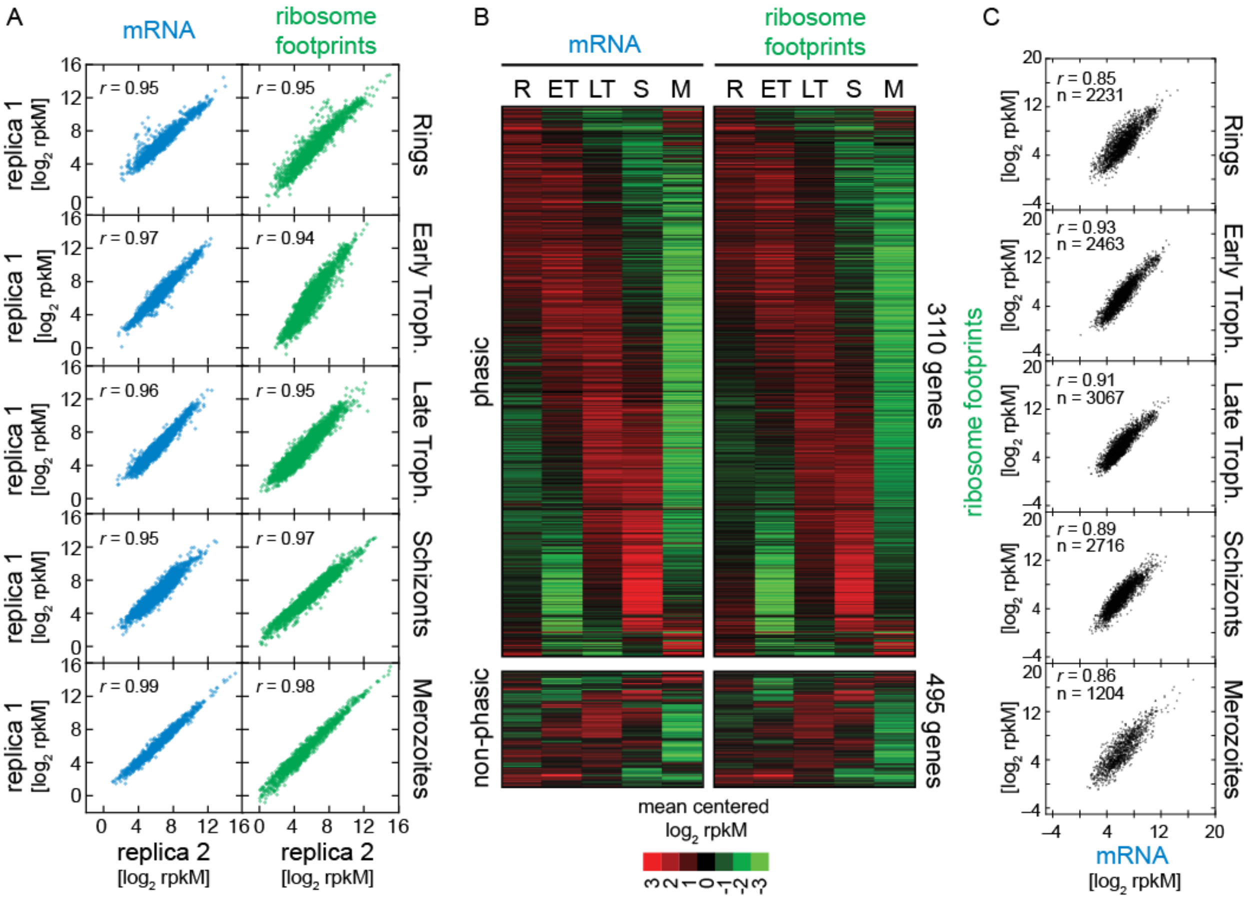


Figure 2. Ribosome profiling through the *P. falciparum* IDC. (A) Reproducibility among biological replicates. Two fully independent biological replicas of each stage were sampled for RNA-seq (left panels, blue) and ribosome profiling (right panels, green). Each dot represents the $\log_2$ rpkM measured for each gene in each stage. *r* = Pearson correlation coefficient. (B) Gene expression and translation are tightly coupled during the *P. falciparum* IDC. Phaseograms of mRNA (left heatmap) and ribosome footprint density (right heatmap) as a function of development for 3110 phasic and 495 non-phasic genes organized in the same order in the left and right heatmap. Data represents mean centered $\log_2$ mRNA and ribosome footprint rpkM values for each gene (rows) in each sampled stage (columns). R = rings, ET = early trophozoites, LT= late trophozoites, S = schizonts, M= merozoites. (C) Log_2 rpkM of mRNA abundance versus ribosome footprint density for all genes expressed (rM ≥32) across the IDC. Pearson correlation coefficients *r* ≥ 85. n = total number of genes.

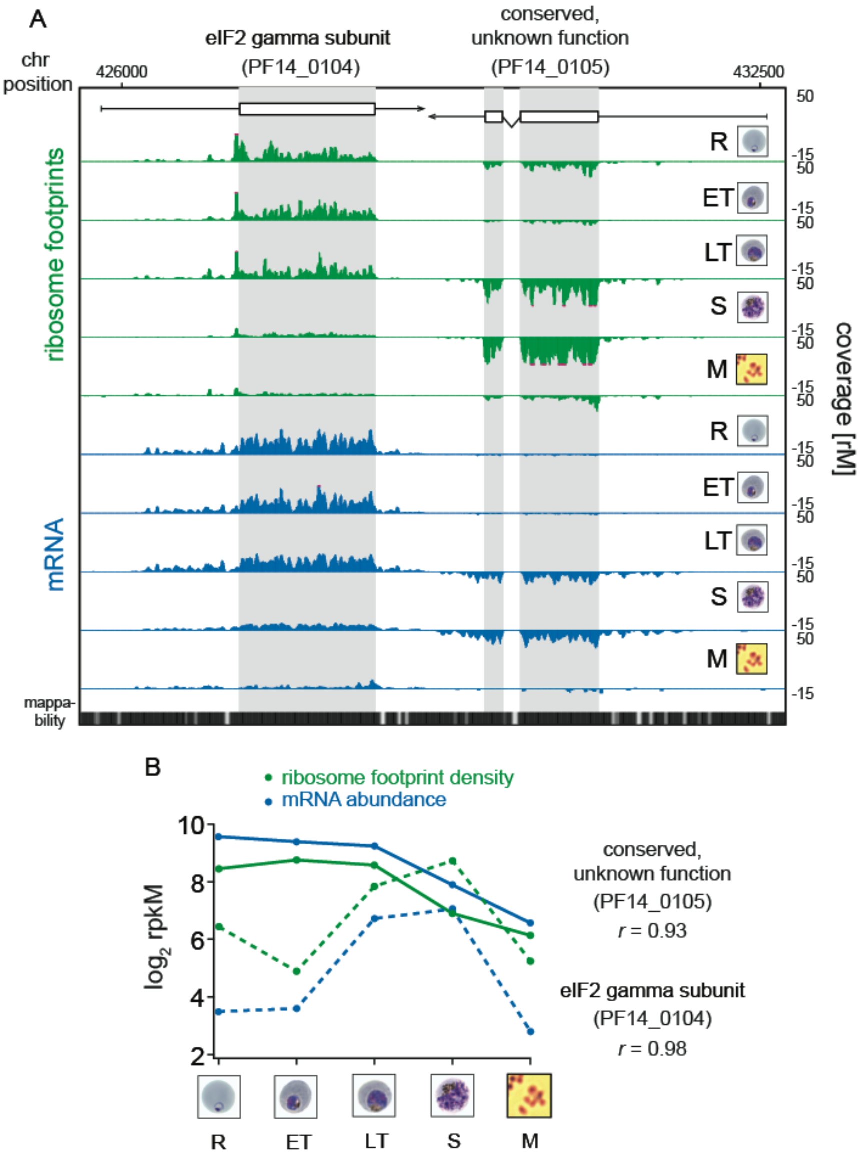


Figure 3. Transcription and translation are highly correlated. (**A**) Ribosome footprint (green) and mRNA (blue) coverage profiles of two neighbor genes, the eIF2 gamma subunit (PF14_0104) and the conserved protein PF14_0105 (CDS, white boxes; HMM-defined UTRs, black lines) in rings (R), early trophozoites (ET), late trophozoites (LT), schizonts (S) and merozoites (M). Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped). (**B**) mRNA and ribosome footprint density of the genes in (A) correlate during development. *r* = Pearson correlation coefficient between ribosome footprint density and mRNA abundance of each gene.

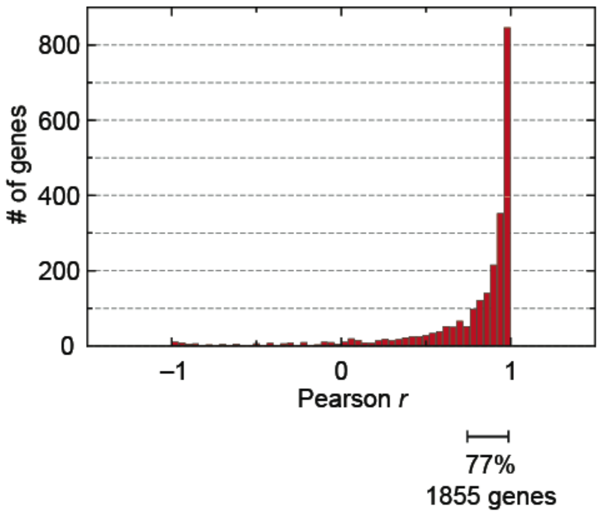


Figure 3-figure supplement 1. mRNA abundance and ribosome footprint density are highly correlated for the majority of genes expressed during the IDC. Pearson correlation of mRNA abundance and ribosome footprint density of every gene expressed in at least 3 stages (2412 genes).

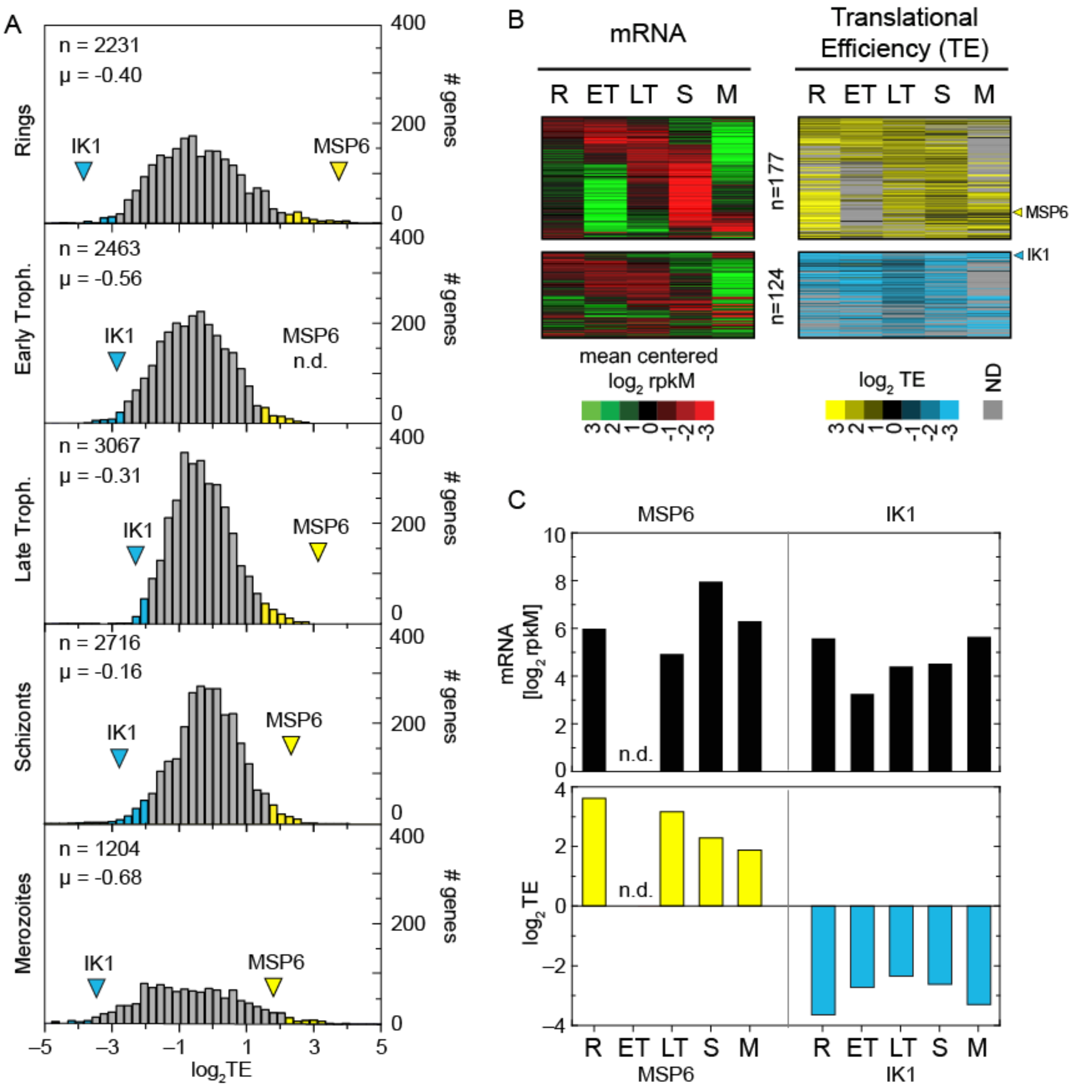


Figure 4. Genome-wide measurements of translation. (**A**) Translational efficiency distributions in each stage. Rings and merozoites have most extreme TE values; ± 2 stdev above (yellow bars) and below (blue bars) the mean. TE values of translationally up-regulated merozoite surface protein (MSP6) and the eukaryotic initiation factor 2 alpha kinase 1 (IK1) (blue arrowhead) across the time course remain high and low, respectively. μ = mean $\log_2$TE, n = total number of genes. (**B**) mRNA abundance and translational efficiency heatmap of translationally up- and down-regulated genes (upper panel and lower panel, respectively). Note TE is independent of changes in mRNA abundance for all genes including MSP6 and IK1 (**C**). R = rings, ET = early trophozoites, LT= late trophozoites, S = schizonts, M= merozoites. n = number of genes, μ = mean, sd. = standard deviation.

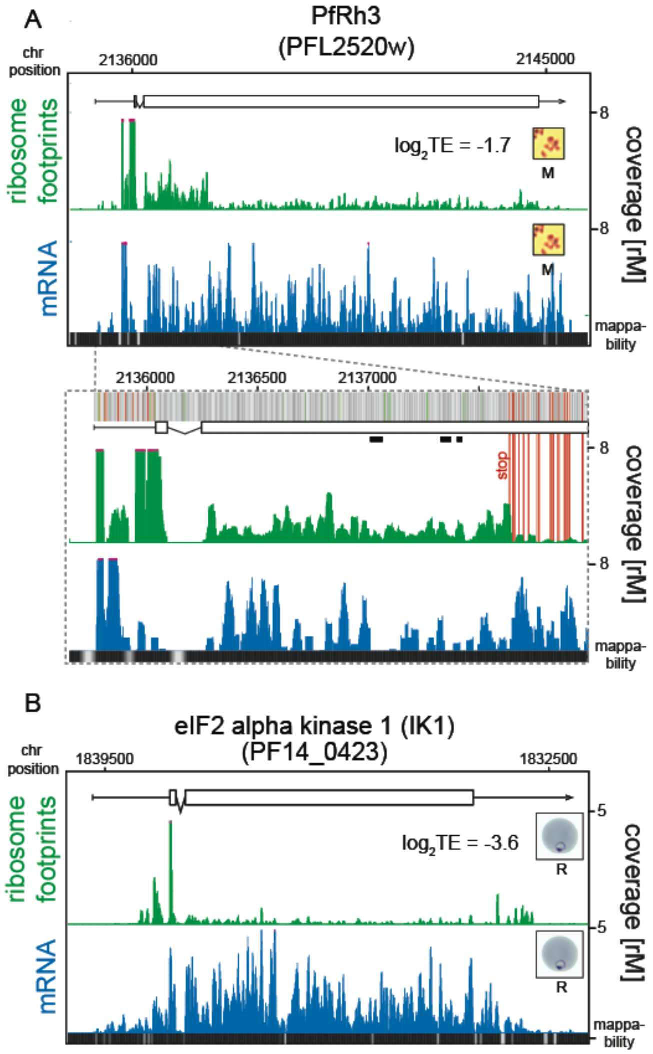


Figure 5. Translationally down-regulated genes have decreased CDS ribosome density. (**A**) Ribosome footprint (green) and mRNA (blue) profiles of the PfRh3 pseudogene (PFL2520w) in merozoites (M). In the detail the bars above the gene model indicate AUG, stop, and any other codon, in green, red, and gray, respectively. Boxes indicate the mapping location of peptides identified by mass spectrometry in gametocytes and sporozoites (Florens et al. 2002; Lasonder et al. 2002). Reduction of ribosome footprint coverage occurs upon encounter of consecutive stop codons (extended red lines). (**B**) eIF2α kinase (PF14_0423) gene in rings (R) showing ribosome footprint accumulation on the 5' leader, 3' UTR and low translational efficiency of the CDS. (CDS, white boxes; HMM-defined UTRs, black lines. Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped).

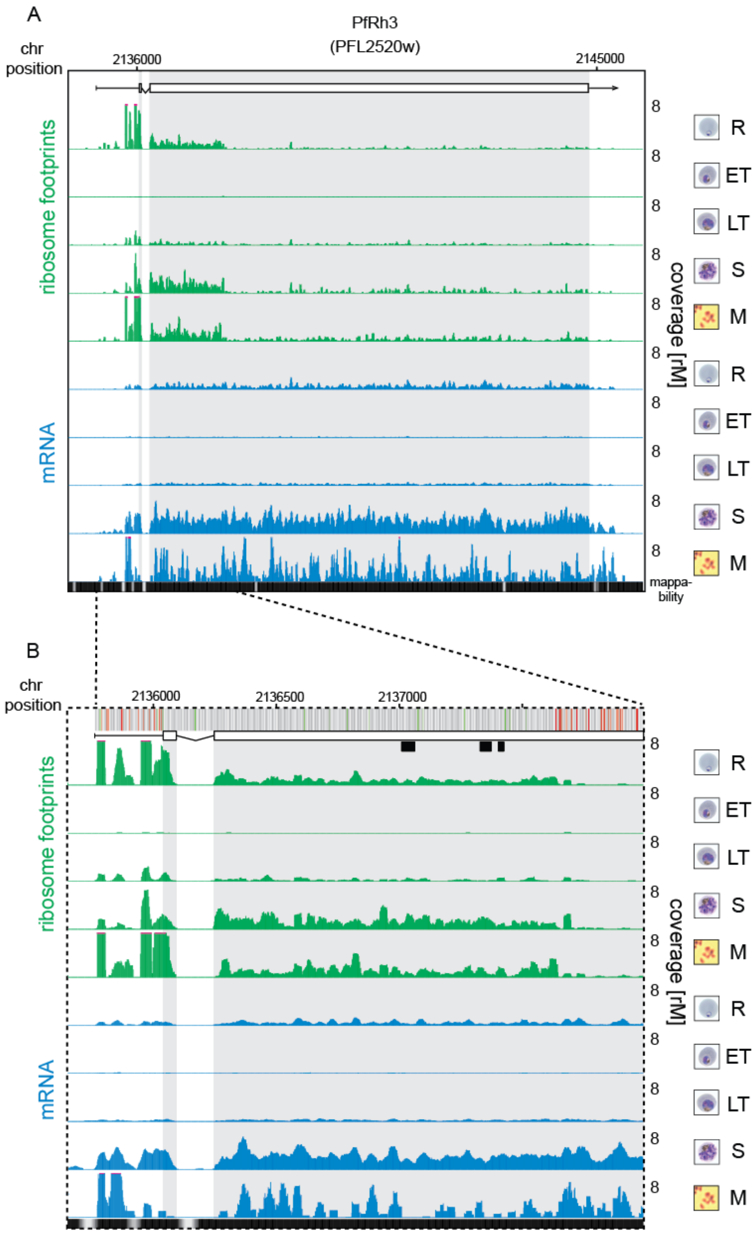


Figure 5-figure supplement 1. Translation of a truncated form of PfRh3 during the IDC. Ribosome footprint (green) and mRNA (blue) profiles of the PfRh3 pseudogene (PFL2520w) in rings (R), early trophozoites (ET), late trophozoites (LT), schizonts (S) and merozoites (M). Translation of PfRh3 occurs until ribosomes dissociate upon the encounter of several consecutive in-frame stop codons. In the detail the bars above the gene model indicate AUG, stop, and any other codon, in green, red, and gray, respectively. Boxes indicate the mapping location of peptides identified by mass spectrometry in gametocytes and sporozoites (Lasonder et al. 2002; Florens et al. 2002).

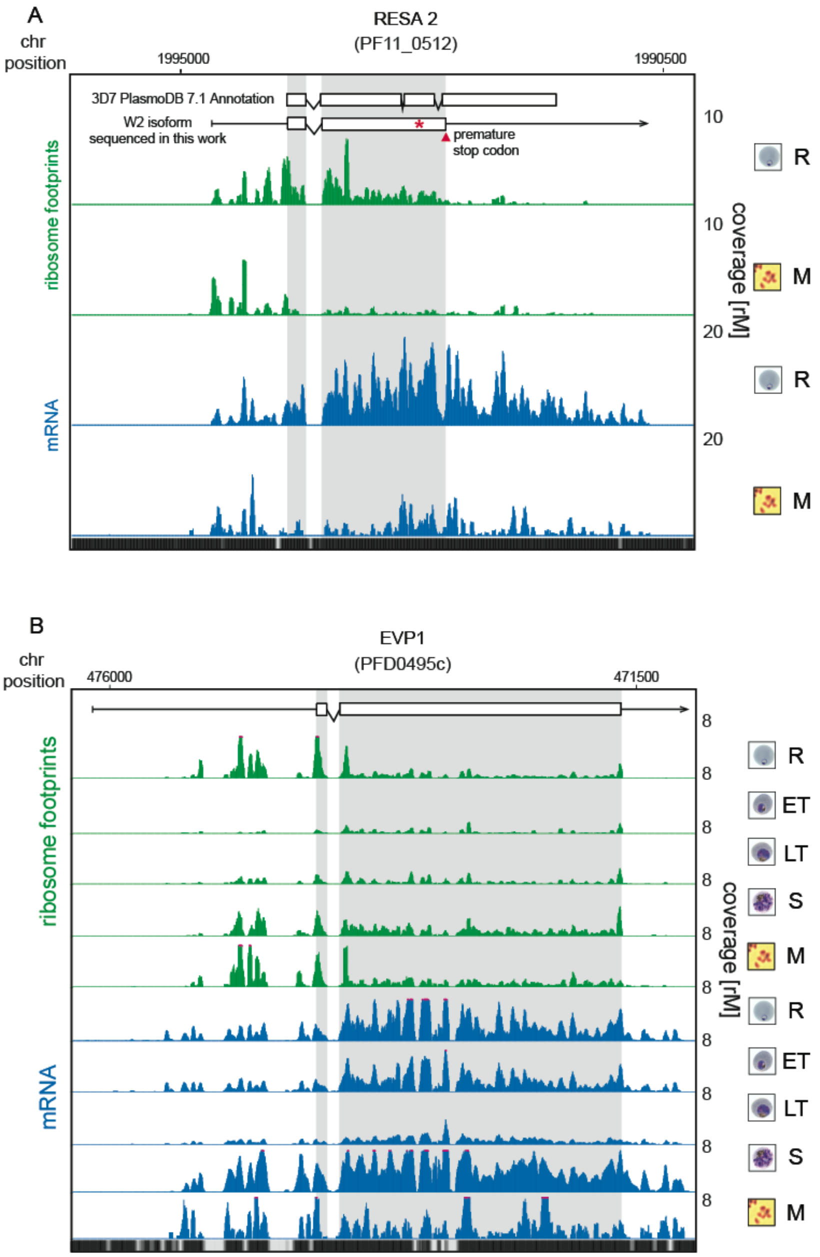


Figure 5-figure supplement 2. Translationally down-regulated genes have decreased CDS ribosome density. (**A**) Ribosome footprint (green) and mRNA (blue) profiles of the ring-infected erythrocyte surface antigen 2, RESA2 pseudogene (PF11_0512) in rings (R) and merozoites (M). Both, the annotated isofoform from PlasmoDB version 7.1 and the gene model for the alternate isoform inferred using ribosome profiling and W2 genomic DNA sequencing data from this study is depicted (CDS, white boxes; HMM-defined UTRs, black lines). The red star indicates a homopolymeric tract in which a single base deletion causes a premature stop codon (red triangle), which coincides with the site of ribosome drop off. (**B**) Ribosome footprint and mRNA profiles of erythrocyte vesicle protein 1, EVP1 (PFD0495c). This gene is transcribed in all stages yet translational efficiencies are relatively low, as evidenced by a depletion of ribosomes on the CDS of the gene particularly in early trophozoites (log$_2$TE = -2.6, -2.9, -1.0, -2.1, -1.4 in rings, early trophozoites, late trophozoites, schizonts and merozoites, respectively). (CDS, white boxes; HMM-defined UTRs, black lines. Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped).

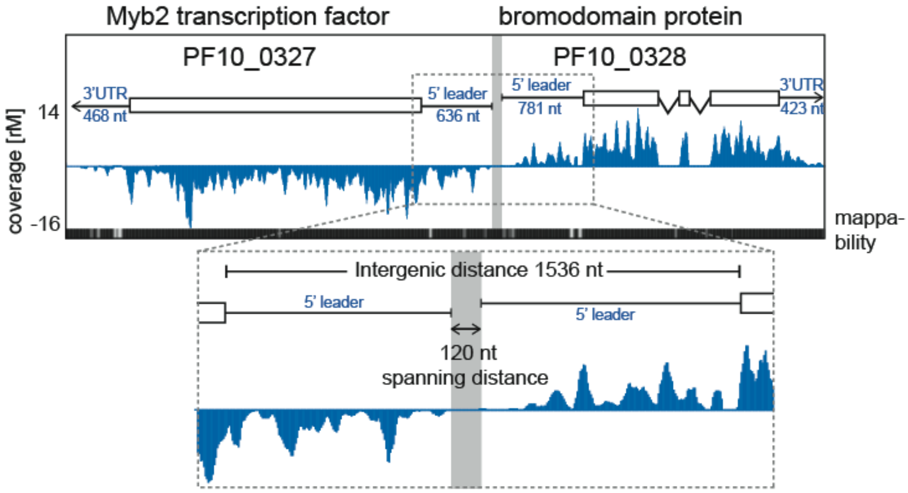
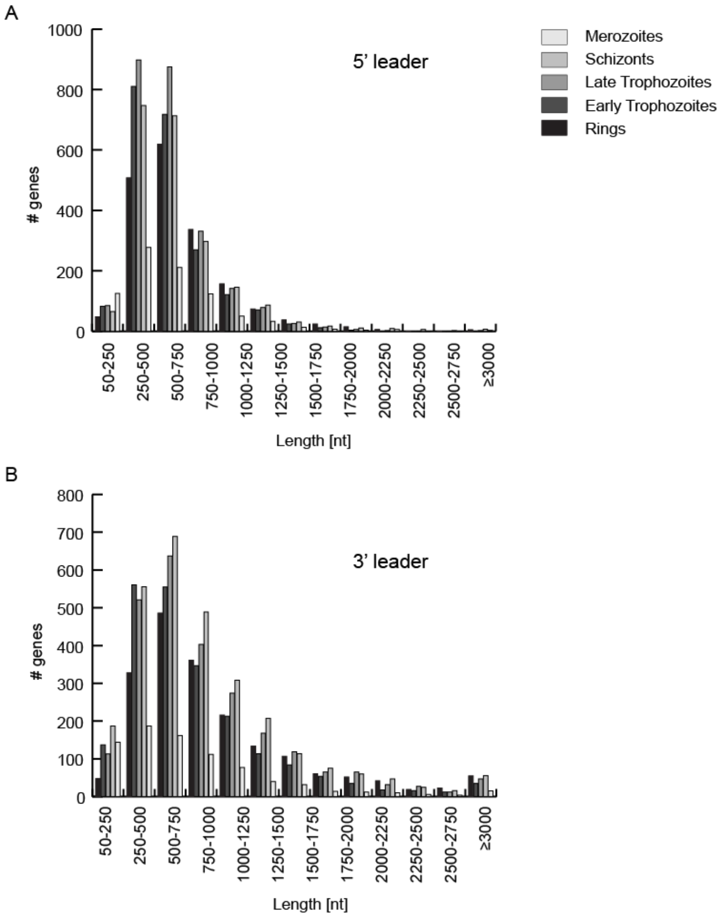


Figure 6. Example of extended transcript annotations using the HMM. 5' leaders and 3' UTRs of the gene pair Myb2 (PF10_0327) and bromodomain protein (PF10_0328) were defined using the HMM designed (see materials and methods). The sizes of 5' leaders and 3' UTRs of these genes in the schizont stage are indicated. The intergenic region is 1536 nt and the spanning distance separating the 5' leaders is 120 nt. Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped).



C

	stage	mean (nt)	median (nt)	stdev	max length (nt)	# genes
5’ Leader	R	1040.4	815	762.9	6810	1934
	ET	852.7	672	674.8	8149	2182
	LT	936.5	736	706.8	8229	2485
	S	927.4	740.5	704.8	7030	2830
	M	787.3	607	716.5	5837	815
3’ UTR	R	724.5	622	407.4	4773	1833
	ET	619.4	546	308.5	2180	2113
	LT	627.2	555	325.0	3195	2462
	S	678.8	572	408.1	4650	2138
	M	634.8	518	585.8	11936	857

Figure 6-figure supplement 1. HMM-defined 5’ leader and 3’ UTR characteristics. 5’ leader (**A**) and 3’ UTR (**B**) length distribution and their statistics (**C**) per stage.

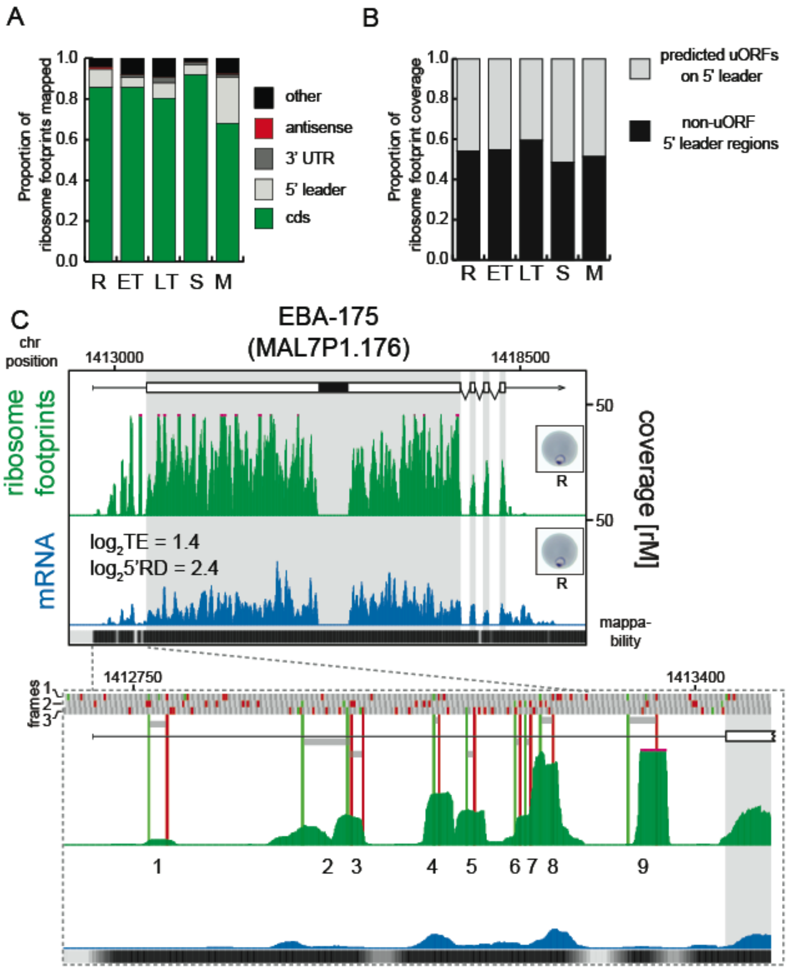
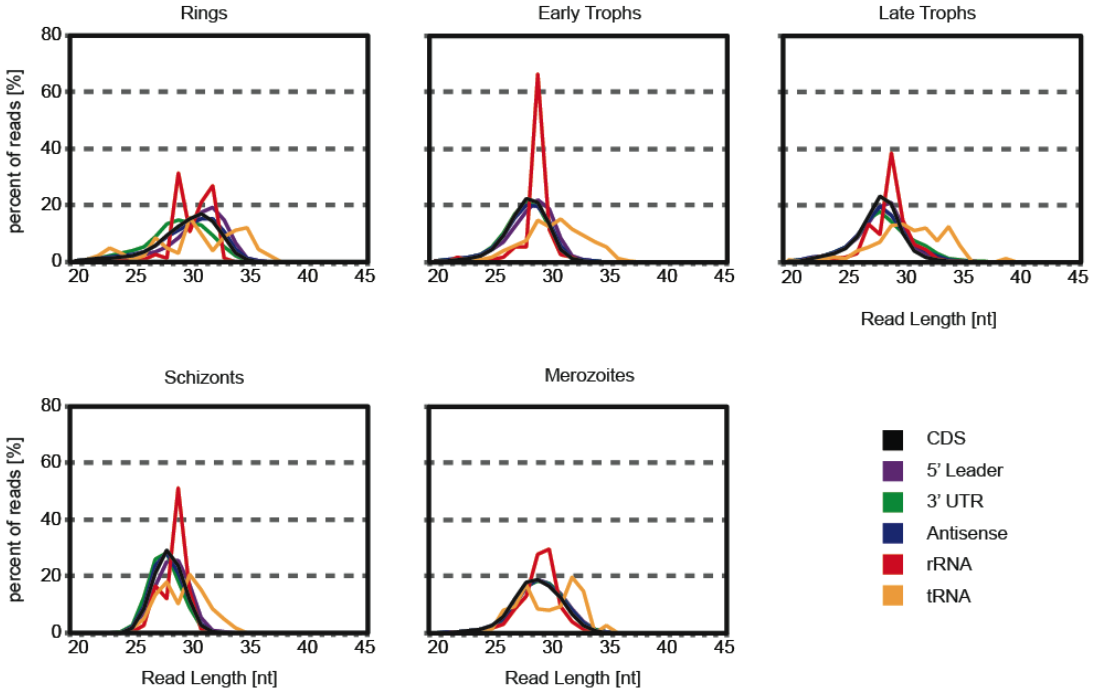


Figure 7. Transcripts accumulate ribosome density within the 5' leader. (**A**) Proportion of mRNA or ribosome footprint reads mapping to CDS, to HMM- defined 5' leaders and 3'UTRs, antisense to annotated coding genes or to other regions of the genome such as mitocondria, plastid, tRNA, rRNA, ncRNA, and 5' leader and 3' UTR regions not defined by the HMM. (**B**) Proportion of ribosome footprints mapping inside or outside predicted uORFs in the HMM- defined 5' leaders. (**C**) Ribosome footprint (green) and mRNA (blue) profiles of the EBA-175 (MAL7P1.176) gene in rings (R) showing ribosome footprint accumulation on the 5' leader. In the detail the bars above the gene model indicate AUG, stop, and any other codon, in green, red, and gray, respectively and in all three possible frames. Gray bars indicate the 9 uORFs present in the 5' leader, starting with an AUG (green line) and ending with a stop codon (red line). Black bar inside CDS indicates a deletion specific to the W2 strain used in this study. CDS, white boxes; HMM-defined UTRs, black lines. Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped).



	KS test pairwise D vs. CDS									
	R		ET		LT		S		M	
	D	p	*D*	p	*D*	p	*D*	p	*D*	p
5' Leader	0.32	6.21x10^{-2}	0.29	1.20x10^{-1}	0.32	6.21x10^{-2}	0.23	3.63x10^{-1}	0.23	3.63x10^{-1}
3' UTR	0.55	8.76x10^{-5}	0.45	2.21x10^{-3}	0.42	5.65x10^{-3}	0.26	2.16x10^{-1}	0.23	3.63x10^{-1}
Antisense	0.61	7.18x10^{-6}	0.52	2.75x10^{-4}	0.58	2.60x10^{-5}	0.35	2.99x10^{-2}	0.39	1.35x10^{-2}
rRNA	0.58	2.60x10^{-5}	0.55	8.76x10^{-5}	0.58	2.60x10^{-5}	0.52	2.75x10^{-4}	0.45	2.21x10^{-3}
tRNA	0.74	2.10x10^{-8}	0.68	4.46x10^{-7}	0.68	4.46x10^{-7}	0.65	1.85x10^{-6}	0.68	4.46x10^{-7}

Figure 7-figure supplement 1. 5' leader footprints are derived from ribosomes. Ribosome footprint read length distributions for reads mapping either to CDSs, 5' leaders, 3' UTRs, antisense, rRNAs or tRNAs are plotted. Read lengths of rRNA and tRNA mapping footprints are significantly different than those mapping to ether the 5' leader, the CDS, or the 3' UTR of transcripts in all stages. KS = Kolmogorov–Smirnov test. D = KS test statistic. R = rings, ET = early trophozoites, LT= late trophozoites, S = schizonts, M= merozoites.

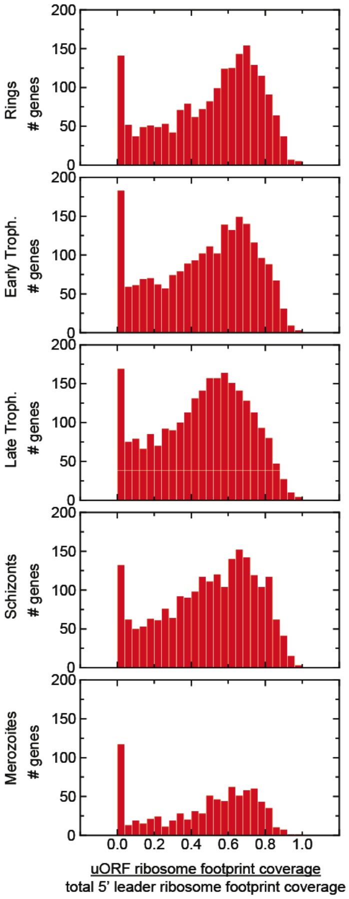


Figure 7-figure supplement 2. Distribution of uORF coverage on 5' leaders of genes expressed during the IDC. The proportion of ribosome footprints mapping inside predicted uORFs was calculated for each gene expressed in each stage. The median of each of these distributions is ~0.5.

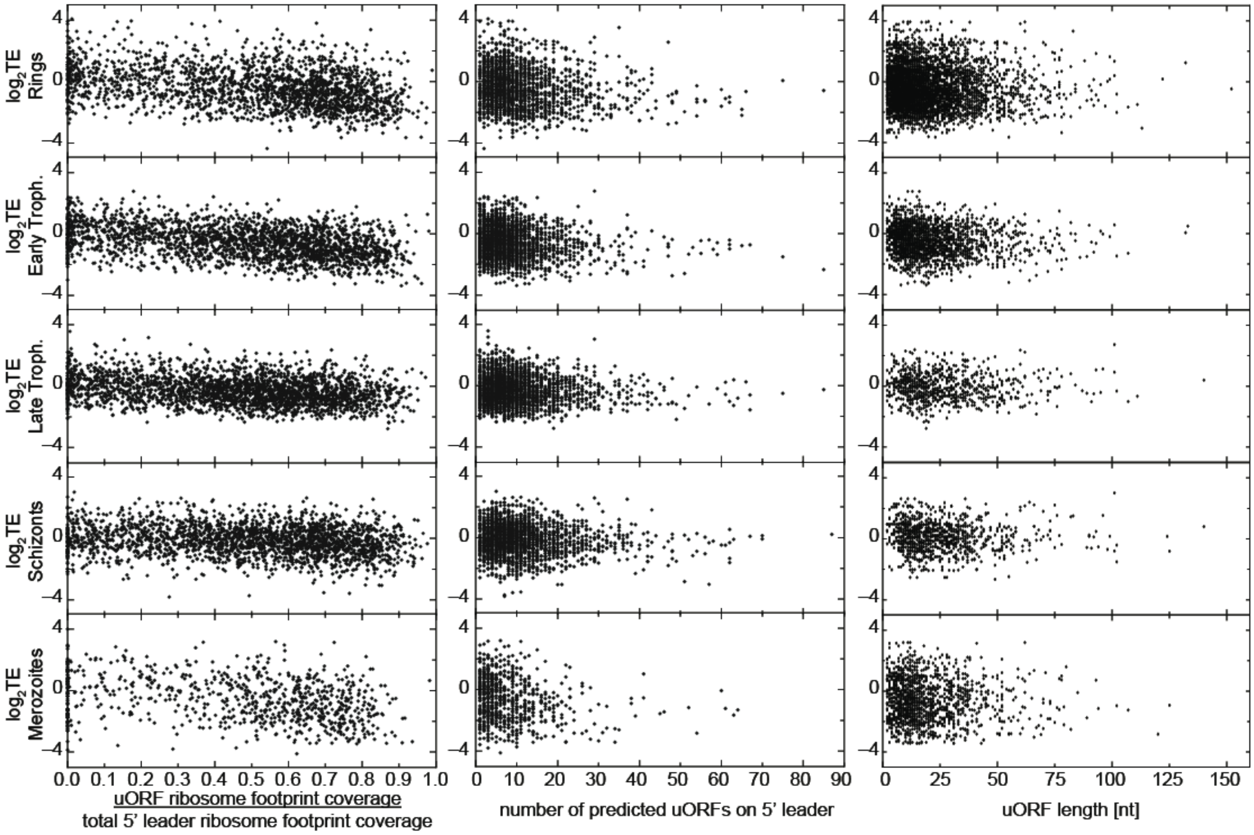


Figure 7-figure supplement 3. uORFs present on 5' leaders have no effect on TE. Translational efficiency (log$_2$TE) for all genes expressed in each stage is plotted against the proportion of reads mapping within uORFs, the number of predicted uORFs, or the length of predicted uORFs in the 5' leader. No direct relationship between these parameters can be observed.

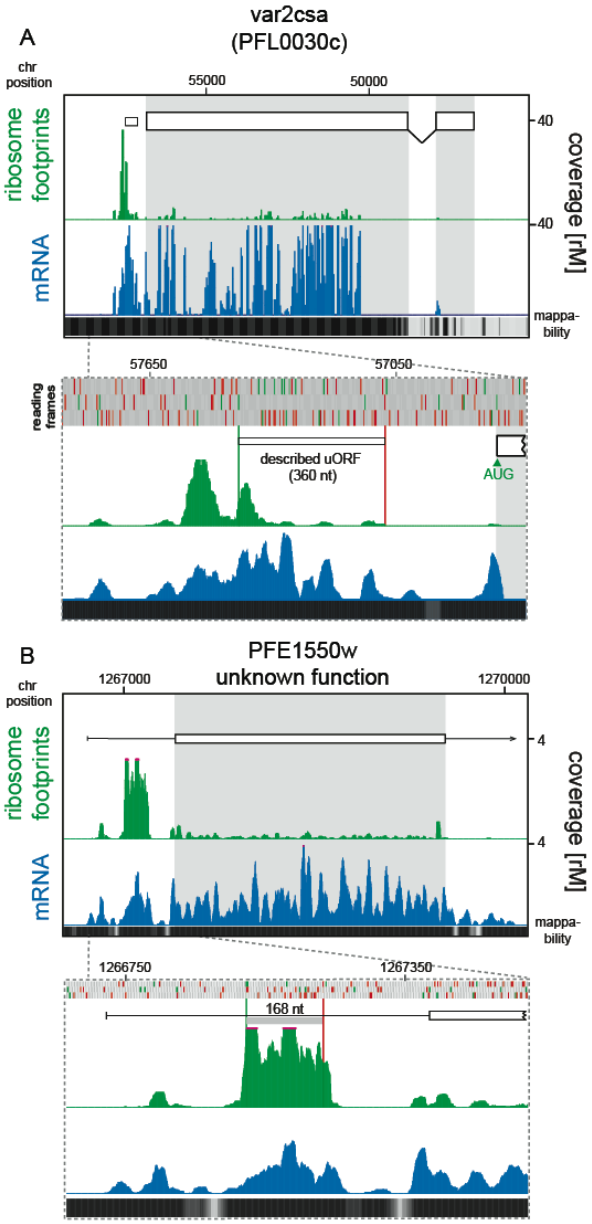


Figure 7-figure supplement 4. Detection of ribosome density on uORFs. (**A**) Ring stage mRNA (blue) and ribosome footprint (green) profiles of VAR2CSA (PFL0030c) are shown. There is virtually no ribosome density on transcript CDS ($\log_2$TERings = -4.2). Ribosomes do accumulate on the previously described (Amulic et al. 2009) 360 nt uORF (white box). This region is depicted in more detail in the panel below where the amino acids, AUGs and stop codons of each of the three reading frames are denoted with gray, green and red bars, respectively. Note that ribosomes start accumulating upstream of the previously described uORF. Mappability at the 3' end of this antigenic variation gene is poor and therefore no mRNA read coverage can be detected here. (**B**) Ring stage mRNA (blue) and ribosome footprint (green) profiles of PFE1550w (unknown function) are shown. Translational efficiency of the CDS is $\log_2 TE_{Rings}$ = -3.6 in rings. 90% of ribosome footprints that map to the 5' leader of this gene accumulate on one of the 6 predicted uORFs (detailed figure below). The predicted uORF is 168 nt (56 aa). Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped).

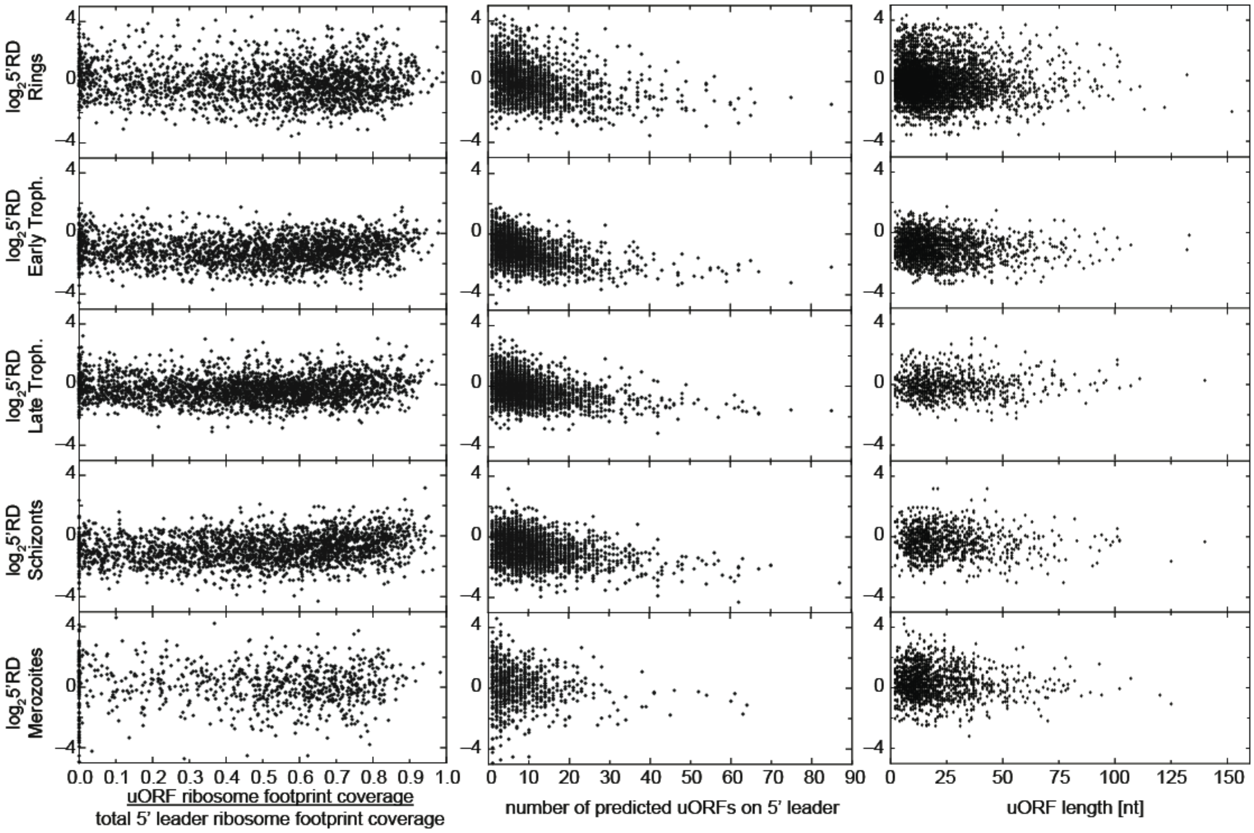


Figure 7-figure supplement 5. uORFs present on 5' leaders have no effect on TE. Ribosome density on the 5' leader (log$_2$5'RD) for all genes expressed in each stage is plotted against the proportion of reads mapping within uORFs, the number of predicted uORFs, or the length of predicted uORFs in the 5' leader. No direct relationship between these parameters can be observed.

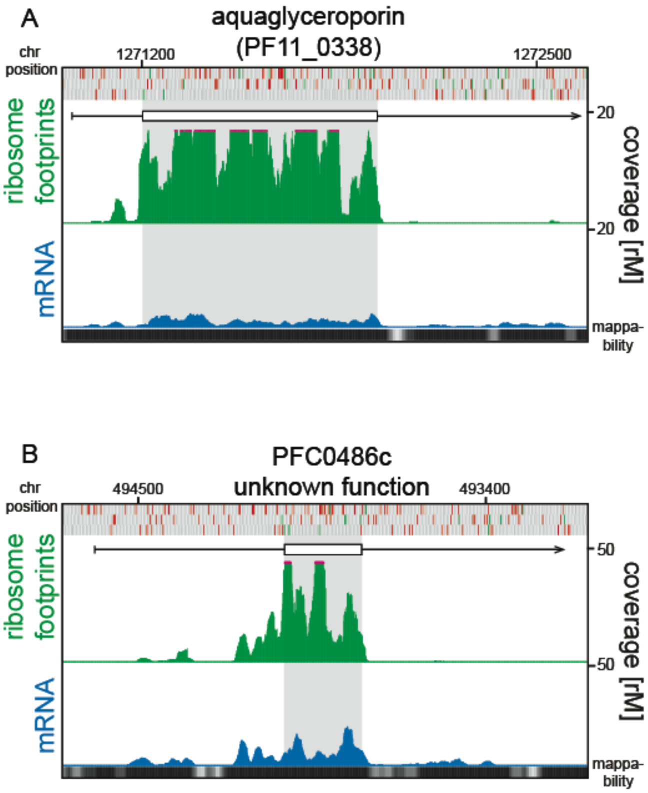


Figure 7-figure supplement 6. 5' ribosome density can be found on 5' leaders devoid of AUGs. Ring stage mRNA (blue) and ribosome footprint (green) profiles of (**A**) aquaglyceroporin (PF11_0338) and (**B**) PFC0486c (unknown function) are shown. Both genes display high ribosome density on their 5' leaders and these are devoid of AUGs. Mappability = mappability score at that position; range 0 (white) to 30 (black). rM = coverage (reads per million reads mapped).

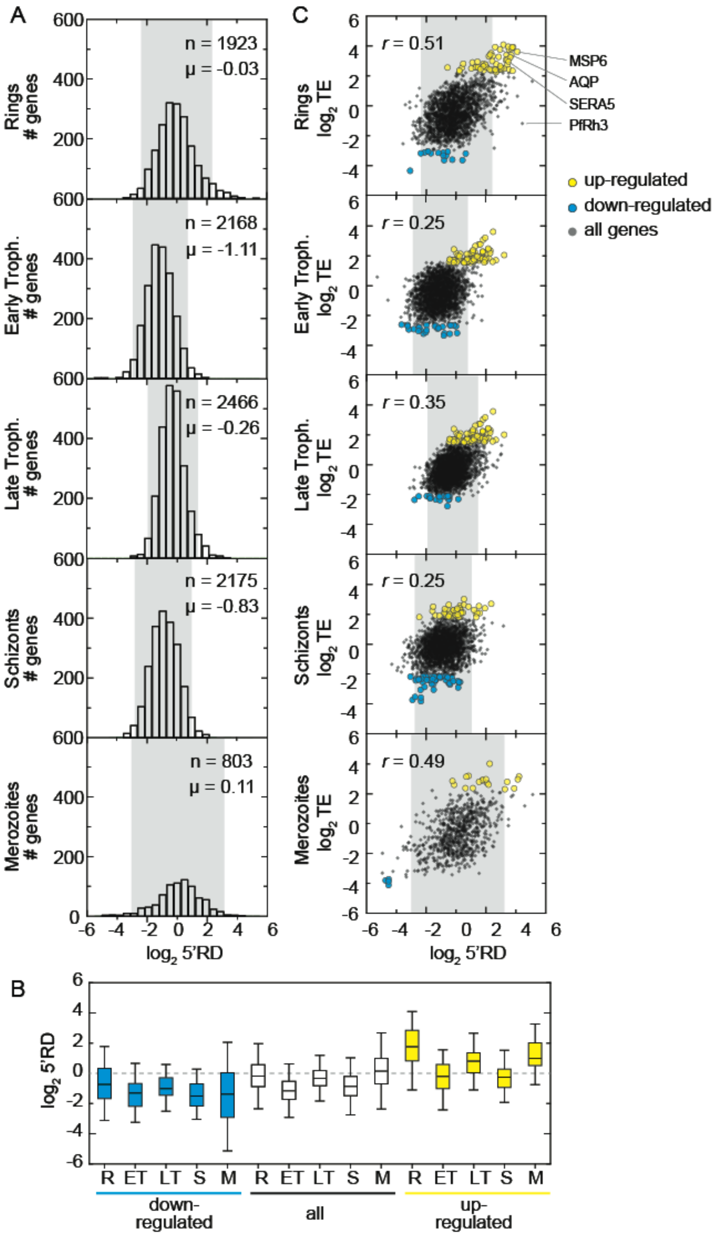


Figure 8. 5' ribosome density is commonly found on genes expressed during the IDC. (**A**) 5' RD distributions in each stage. Transcripts in rings and merozoites have on average higher 5' RD values; ± 2 stdev values lie outside gray shade. μ = mean log$_2$5'RD, n = total number of genes. (**B**) 5'RD values of the translationally up-regulated set of genes (yellow boxes) relatively higher (average log$_2$5'RD R = 1.73, ET = -0.26, LT = 0.78, S = 0.30, M = 1.16.) than the rest (white boxes) or the set of down-regulated (blue buxes) genes. (**C**) 5'RD weakly correlates with translational efficiency. The translationally up-regulated gene set (yellow circles) is associated with high 5' RD, particularly in rings. The translationally up-regulated genes merozoite surface protein (MSP6), aquaglyceroporin (AQP), serine repeat antigen (SERA5), and the reticulocyte binding protein homologue 3 (PfRh3) are pointed out. *r* = Pearson correlation coefficient.

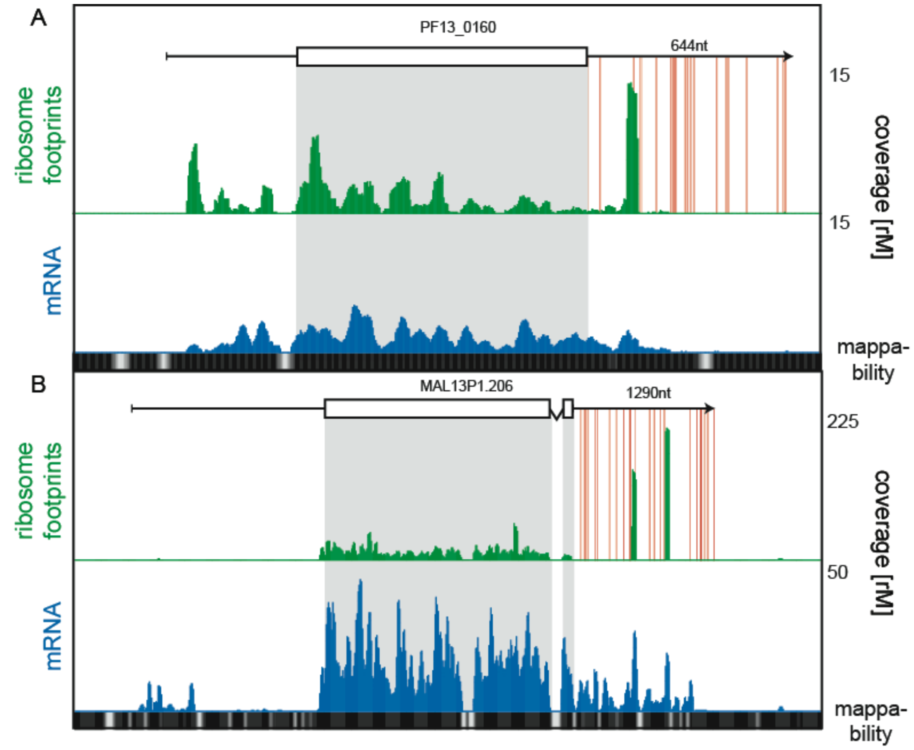


Figure 9. 3'UTR ribosome density. (**A**) Late trophozoite stage mRNA (blue) and ribosome footprint (green) profiles of the conserved plasmodium protein, PF13_0160. Ribosomes can be detected up to ~130 nt beyond the stop codon on the 3'UTR and accumulate in a single large peak. Red lines indicate in-frame stop codons on the 3' UTR. (**B**) Two large peaks of ribosome footprint density can be detected 560 nt and 860 nt downstream from the stop codon in the 3'UTR of the sodium-dependent phosphate transporter, MAL13P1.206.

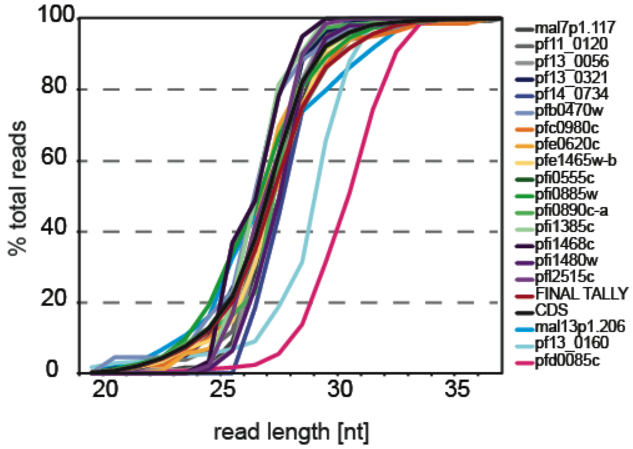


Figure 9-figure supplement 1. 3'UTR ribosome footprint size distribution. Cummulative read length distributions of all reads mapping to the 3' UTR of the 19 genes with 3' ribosome density identified compared to the read length distributions of reads mapping to all CDSs in the late trophozoite stage (black line). Footprint length distributions for MAL13P1.206, PF13_0160 and PFD0085c are least similar to the ribosome footprints that map to the CDS.

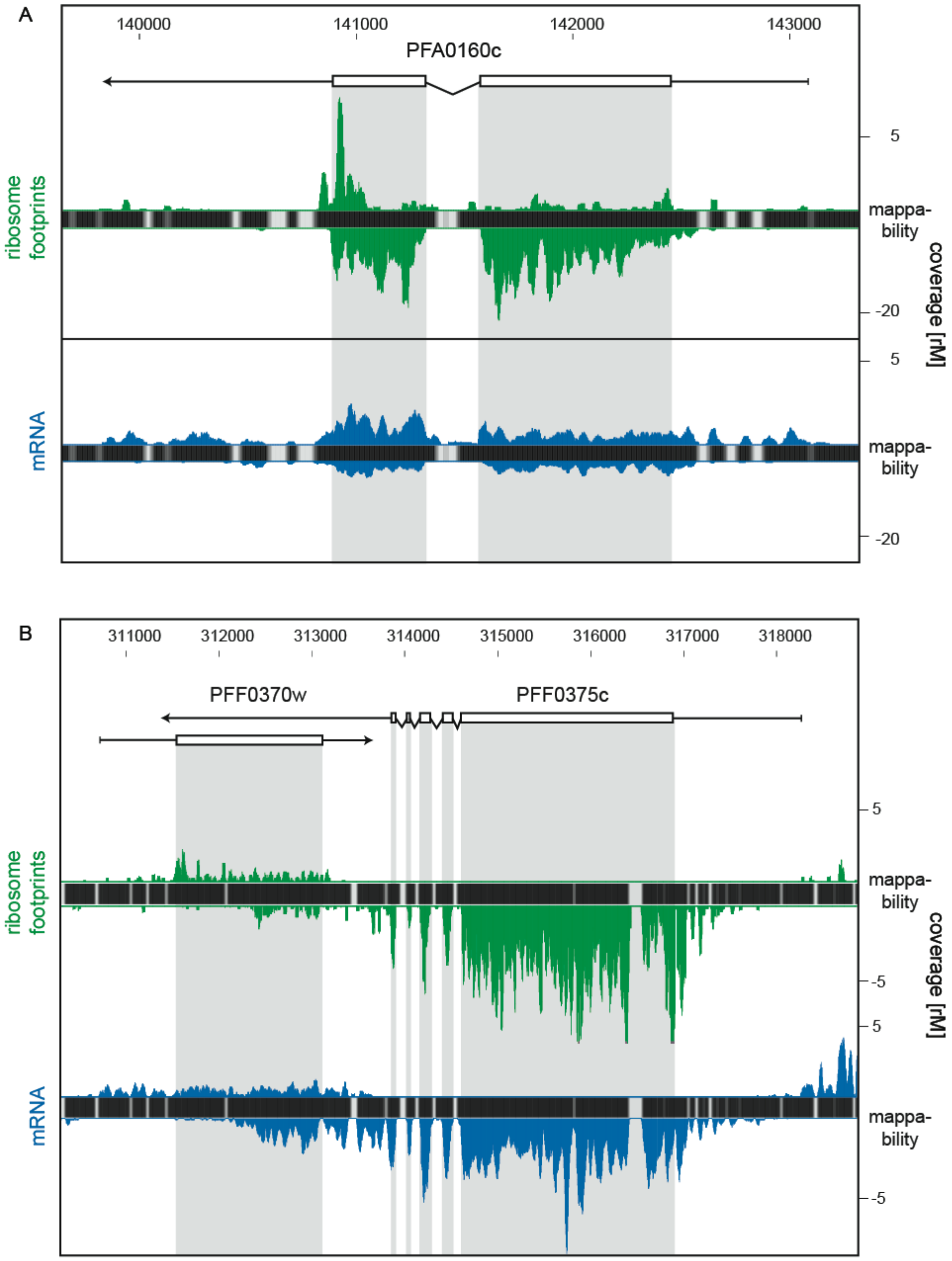


Figure 10. Strand specific libraries can distinguish antisense from sense gene transcription. (**A**) Schizont stage mRNA (blue) and ribosome footprint (green) profiles of the nucleoside transporter pfNT4 (PFA0160c). The antisense transcript covers the full extent of the sense transcript and displays ribosome density. (**B**) An example of antisense reads originating from a neighboring UTR in the schizont stage. The antisense reads in the para-hydroxybenzoate polyprenyltransferase (PFF0370w) stem from the 3' UTR of the neighboring conserved plasmodium protein (PFF0375c).

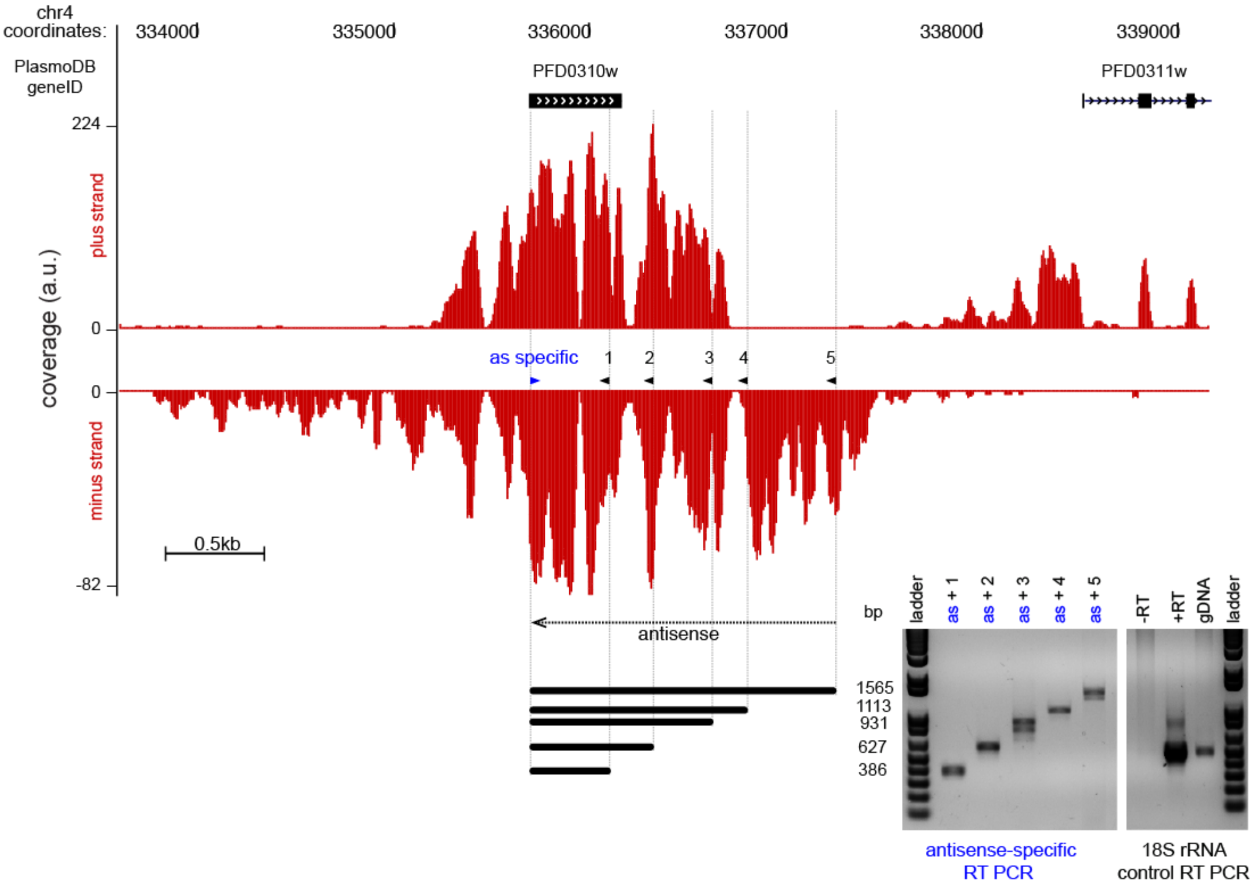


Figure 10-figure supplement 1. Strand-specific RT-PCR detection of the antisense to Pfs16. Read coverage on the plus and minus strand of the stage-specific protein precursor Pfs16 (PFD0310w) locus. The gene is encoded on the plus strand and the antisense transcript covers and extends beyond the sense transcript ~3.7kb. The strand-specific primer was used for both reverse transcription and as forward primer for the PCR (blue arrowhead). The 5 PCR primers (black arrowhead) and the expected amplicon sizes are shown next to the strand-specific RT-PCR results. 18S rRNA primers were used in the control reactions. a.u. = arbitrary units

Chapter 5: Development of an *in vitro* translation assay and a screen to discover inhibitors of protein synthesis

This chapter is a summary of work done by:

Ahyong V, Leon K, Witchley J, Ebert D, DeRisi JL.

Author contributions:

Vida Ahyong developed the *in vitro* translation assay with the help of Jessica Witchley and Krisoffer Leon. Screening of the MMV malaria box was done by Vida Ahyong and Kristoffer Leon. Joseph L. DeRisi conceived and supervised the project.

Introduction

Identifying new antimalarials with novel mechanisms of action is a key goal in the fight to eradicate malaria worldwide. Many strategies rely on screening in vitro cultures of Plasmodium falciparum against large compound collections whose mechanisms of action are unknown and assaying for growth inhibition in a “top-down” approach to drug discovery. Previously, our lab has focused on this type of approach and the subsequent identification of drug targets are discovered by selection of resistant strains and whole genome sequencing of these resistant strains to identify mutations (i.e.: Single nucleotide polymorphisms, copy number variants, insertions/deletions) that confer resistance[1][2][3]. However, one challenge encountered in the discovery of novel antimalarial compounds is the possibility that these new drugs will be cross-resistant with previously identified antimalarial compounds because they share the same drug target. This has been the case for the drug target PfATP4 in which multiple groups have identified diverse chemical compounds that all have the same target and quite possibly the same mechanism of action[4][5]. The consequence of discovering compounds with the same overlapping target is a narrower panel of potential drug therapies when drug resistance inevitably emerges.

An alternate approach is to assay for the activity against a specific biological function, resulting in hits with a known target or pathway, in a “bottom-up” approach to drug discovery. The advantages of a target-based approach are that it involves an understanding of a selected gene or pathway in order to develop a new drug candidate. The recent release of the Medicines for Malaria Venture (MMV) Malaria Box is a collaborative effort between St. Judes, GSK, and Novartis to give the malaria research community access to new compounds with undescribed targets and mechanisms of action that provides a great opportunity for target-based screens of P. falciparum[6]. The library contains 400 chemically diverse compounds that are commercially available and pre-screened for activity in the blood stages of *Plasmodium falciparum* with minimal cytotoxicity.

Protein synthesis inhibitors are common in antibiotics because they take advantage of the large differences between prokaryotic and eukaryotic ribosomes. Despite *Plasmodium falciparum* being a eukaryotic organism, there are ample differences between it and the mammalian ribosome[7][8]. We hypothesize that these sequence and structural differences will result in enough differences to chemically disrupt the activity of the Pf 80S ribosome but not the mammalian ribosome. Recently, a potent new compound, DDD107498, was reported to specifically inhibit P. falciparum protein synthesis by blocking activity of the Plasmodium falciparum translation elongation factor 2 (eEF2)[9]. Here we report a high throughput in vitro translation assay to discover specific inhibitors of the malaria ribosome present in the Malaria Box. The ultimate goal is to discover compounds with new and separate mechanisms of action to prevent global resistance to the current rostrum of clinically available antimalarials.

Results

Development of a high-throughput malaria specific in vitro translation assay

Building upon the work of Ferreras et al (2000) we further optimized and developed an *in vitro* translation assay prepared from *Plasmodium falciparum* cultures with the addition of exogenous T7 luciferase reporter mRNAs to allow for high-throughput plate based luciferase assay screening (Figure 1A)[10]. We scaled our cultures up to 500ml hyperflasks (Corning) using synchronized, high-density late trophozoite cultures followed by a saponin lysis to release the parasites from the red blood cells (see Materials and Methods). Harvesting higher parasitemia cultures resulted in robust and reproducible translation activity when the A_{280} measured in the range of 7-10mg/mL whereas we found that no viable conditions when spin based concentrator columns were used to concentrate the lysate post-harvest. We tested multiple lysis conditions (data not shown) including mechanical shearing, freeze/thawing cycles, hypotonic lysis and found the most success using a cell homogenizer (Isobiotec, Germany), which gently and uniformly breaks open cells by passing the lysate though a precise 4um ball bearing clearance, a technique that has found widespread use in subcellular fractionation studies[11]. We routinely obtain cell-free lysate that is competent for *in vitro* translation as measured by the translation of the firefly luciferase reporter constructs with a *Plasmodium falciparum* specific 5' untranslated region (UTR), of the ubiquitously expressed erythrocyte binding antigen, EBA-175 (Figure 1B). These extracts achieve luciferase saturation similar to commercially available rabbit reticulocyte lysate (Retic IVT, Ambion), though with slower kinetics presumably due to a higher protein concentration of rabbit reticulocyte lysate with a measured A_{280} ~ 40mg/mL compared to the 7-10mg/mL concentration of *P. falciparum* lysates. Each *P. falciparum* lysate performs slightly different in terms of the time to reach its maximum luciferase output and the maximal expression, thus for each harvest, a preliminary kinetic assay is necessary to establish the final incubation time.

To test whether this assay can discover inhibitors specific to protein synthesis, we compared the translation activity of our lysate in the presence of 5uM cycloheximide (33x reported IC_{50}=150nM), an inhibitor of general eukaryotic ribosome elongation. When cycloheximide is added prior to the 37°C 120' incubation, translation is inhibited resulting in the absence of luciferase expression (Figure 1A). Furthermore, we want to determine that our assay is specific for inhibitors of protein synthesis and not general antimalarial activity so we tested several antimalarial compounds at 1uM concentration (dihydroartemisinin, emetine, piperaquine, chloroquine, SJ579, Quinine, and Primiquine) for their effect on *P. falciparum* translation (Figure 2A). The only compound that showed a considerable decrease in translation was emetine which had ~50% inhibition though not to the same extent as cycloheximide. Emetine has previously been identified as a anti-protozoal compound and an inhibitor of protein synthesis in eukaryotes[12][13]. The surprising result is that emetine did not inhibit protein synthesis to the extent that we expected with only a 50% inhibition at 1uM drug concentration yet the reported EC50 for emetine is 10nM[14]. Follow up inhibitory concentration curves will determine the IC50 concentration of emetine is in fact much higher than the reported EC50.

Screening the open access Malaria Box for translational inhibitors

We arbitrarily chose to screen the malaria box at a 1 μM drug concentration end point assay for 1.5 h, or before the assay reaches saturation (~80% of maximum translation per lysate). Each plate was replicated at least three times and normalized by the average of all the control wells situated at the peripheral columns of each plate. Due to the extended temporal nature of the luciferase measurements on the luminometer, we noticed a positive correlation between time of the luciferase assay and the position on the 96 well assay plate. To negate these changes of translation during the assay, we added a 5uM cycloheximide stop solution to each well to stop all translation at the same time providing for uniform temporal measurements of translation throughout the luminescence assay. The results of this screen gave a fractional value of the maximum *in vitro* translation normalized to the average of no drug controls (Figure 2B). Further hits were prioritized by the correspondence between replicates whereby both replicates gave at least 20% translation inhibition and a standard deviation less than 20% to pass our first filter (Figure 3A). Fifteen compounds exhibited this minimum threshold of inhibition (Table 1). Though the malaria box has previously been screened for cytotoxic effects, to ensure that inhibitors did not generally block eukaryotic *in vitro* translation or were inhibitors of luciferase we performed a secondary screen using commercially available rabbit reticulocyte lysate (Retic IVT, Life Tech) against the 15 hits, also in 1uM drug concentration with 2 replicates. We define a specificity index as the ratio of normalized *P. falciparum* average translation over the normalized rabbit reticulocyte average translation. Specificity ratios ranged from 0.48 to 2.35 where the low values signify compounds that have a greater effect on *P. falciparum* translation whereas higher values signify compounds that have a lesser effect on *P. falciparum* translation compared to a rabbit reticulocyte system, a proxy for general eukaryotic translation. Eleven of these compounds showed a ratio of <0.9. Future studies will profile the inhibitory concentration (IC) curves to determine the IC_{50} value of lead compounds with this assay.

Discussion

Here we present a novel high throughput luciferase assay that allows for the discovery of compounds inhibiting protein synthesis in *P. falciparum*. We show that this cell-free system is sensitive to known translation inhibitors, cycloheximide and emetine, whereas antimalarials with no known effect on translation such as chloroquine and dihydroartemisinin, do not show an *in vitro* inhibitory effect compared to their known effect on *in vivo* cultures of *P. falciparum*. This study provides a list of 12 compounds that have a specific effect (specificity index <0.9) on *P. falciparum* translation that will be further validated and structure activity relationship determined in future studies for their mechanism of inhibition on the *P. falciparum* ribosome. Though our screening of the Malaria Box did not result in strong malaria-specific inhibitors of translation, the assay developed here will be a powerful tool for additional drug screens and validation of drugs with proposed mechanisms of action affecting *P. falciparum* translation.

Materials and Methods

Plasmodium falciparum culturing

Strains of W2 were maintained in Hyperflasks (Corning, Corning, NY) in 500 mL RPMIc (RPMI 1640 media supplemented with 0.25% Albumax II (GIBCO, Grand Island, NY), 2g/L sodium bicarbonate, 25mM HEPES (pH 7.4), 0.1 mM hypoxanthine, and 50 ug/L gentamycin) in a 37°C, 5% O2, 5% CO2 incubator in 2% hematocrit (HC). Cells were synchronized with 5% sorbitol treatment for 2 generations to achieve high synchronicity.

Harvesting cell pellets

Parasite cultures were harvested at the late trophozoite stage to approximately 15% parasitemia by centrifugation for 5 min at 1500xg at room temperature and 0.1% final saponin in Buffer A (20 mM HEPES pH 8.0, 2 mM Mg(OAc)2, 120 mM KOAc). Saponin lysed pellets were centrifuged at 4°C 10,000xg for 10 min and washed once with ice cold Buffer A. The pellet was resuspended in 2 mL of Buffer B2 (20 mM HEPES pH 8.0, 100 mM KOAc, 0.75 mM Mg(OAC)2, 2mM DTT, 20% glycerol), flash frozen, and stored in -80°C freezer until the sample was ready to homogenize.

Homogenization of cell pellets

Frozen pellets were thawed on ice and added to a 3 mL luer lock syringe, locked onto a pre-chilled cell homogenizer (Isobiotec, Germany) on ice and passed between two syringes 20 times. Lysate was centrifuged at 4°C 16,000xg for 10 minutes and the supernatant was saved at -80°C until ready for the assays.

In vitro translation assay

In vitro translation reaction were carried out with the following components in 20 uL: 16 uL lysate, 1 ug T7 transcribed firefly luciferase mRNA, 10 μM amino acid mixture, 20 mM HEPES/KoH pH 8.0, 75 mM KoAc, 2 mM Mg(OAc)2, 2 mM DTT, 0.5 mM ATP, 0.1 mM GTP, 20 mM creatine phosphate, 0.2 ug/ul creatine kinase for 1.5h at 37°C. After incubation, the reactions were quenched with 5 μM cycloheximide. Reactions were assayed using the Promega GloMax-Multi+ microplate reader with a 3 second delay and 10 second integration after addition of luciferin reagent.

References

1. Jiménez-Díaz MB, Ebert D, Salinas Y, Pradhan A, Lehane AM, Myrand-Lapierre M-E, O'Loughlin KG, Shackleford DM, Justino de Almeida M, Carrillo AK, Clark JA, Dennis ASM, Diep J, Deng X, Duffy S, Endsley AN, Fedewa G, Guiguemde WA, Gómez MG, Holbrook G, Horst J, Kim CC, Liu J, Lee MCS, Matheny A, Martínez MS, Miller G, Rodríguez-Alejandre A, Sanz L, Sigal M, Spillman NJ, Stein PD, Wang Z, Zhu F, Waterson D, Knapp S, Shelat A, Avery VM, Fidock DA, Gamo F-J, Charman SA, Mirsalis JC, Ma H, Ferrer S, Kirk K, Angulo-Barturen I, Kyle DE, DeRisi JL, Floyd DM, Guy RK. (+)-SJ733, a clinical candidate for malaria that acts through ATP4 to induce rapid host-mediated clearance of Plasmodium. Proc Natl Acad Sci USA. 2014 Dec 16;111(50):E5455–5462.

2. Guler JL, Freeman DL, Ahyong V, Patrapuvich R, White J, Gujjar R, Phillips MA, DeRisi J, Rathod PK. Asexual populations of the human malaria parasite, Plasmodium falciparum, use a two-step genomic strategy to acquire accurate, beneficial DNA amplifications. PLoS Pathog. 2013;9(5):e1003375.

3. Wu W, Herrera Z, Ebert D, Baska K, Cho SH, DeRisi JL, Yeh E. A chemical rescue screen identifies a Plasmodium falciparum apicoplast inhibitor targeting MEP isoprenoid precursor biosynthesis. Antimicrob Agents Chemother. 2015 Jan;59(1):356–364.

4. Lehane AM, Ridgway MC, Baker E, Kirk K. Diverse chemotypes disrupt ion homeostasis in the Malaria parasite. Mol Microbiol. 2014 Oct;94(2):327–339.

5. Flannery EL, McNamara CW, Kim SW, Kato TS, Li F, Teng CH, Gagaring K, Manary MJ, Barboa R, Meister S, Kuhen K, Vinetz JM, Chatterjee AK, Winzeler EA. Mutations in the P-type cation-transporter ATPase 4, PfATP4, mediate resistance to both aminopyrazole and spiroindolone antimalarials. ACS Chem Biol. 2015 Feb 20;10(2):413–420.

6. Spangenberg T, Burrows JN, Kowalczyk P, McDonald S, Wells TNC, Willis P. The open access malaria box: a drug discovery catalyst for neglected diseases. PLoS ONE. 2013;8(6):e62906.

7. Li J, McConkey GA, Rogers MJ, Waters AP, McCutchan TR. Plasmodium: the developmentally regulated ribosome. Exp Parasitol. 1994 Jun;78(4):437–441.

8. Jackson KE, Habib S, Frugier M, Hoen R, Khan S, Pham JS, Ribas de Pouplana L, Royo M, Santos MAS, Sharma A, Ralph SA. Protein translation in Plasmodium parasites. Trends Parasitol. 2011 Oct;27(10):467–476.

9. Baragaña B, Hallyburton I, Lee MCS, Norcross NR, Grimaldi R, Otto TD, Proto WR, Blagborough AM, Meister S, Wirjanata G, Ruecker A, Upton LM, Abraham TS, Almeida MJ, Pradhan A, Porzelle A, Martínez MS, Bolscher JM, Woodland A, Norval S, Zuccotto F, Thomas J, Simeons F, Stojanovski L, Osuna-Cabello M, Brock PM, Churcher TS, Sala KA, Zakutansky SE, Jiménez-Díaz MB, Sanz LM, Riley J, Basak R, Campbell M, Avery VM, Sauerwein RW, Dechering KJ, Noviyanti R, Campo B, Frearson JA, Angulo-Barturen I, Ferrer-Bazaga S, Gamo FJ, Wyatt PG, Leroy D, Siegl P, Delves MJ, Kyle DE, Wittlin S, Marfurt J, Price RN, Sinden RE, Winzeler EA, Charman SA, Bebrevska L, Gray DW, Campbell S, Fairlamb AH,

Willis PA, Rayner JC, Fidock DA, Read KD, Gilbert IH. A novel multiple-stage antimalarial agent that inhibits protein synthesis. Nature. 2015 Jun 17;522(7556):315–320.

10. Ferreras A, Triana L, Correia H, Sánchez E, Herrera F. An in vitro system from Plasmodium falciparum active in endogenous mRNA translation. Mem Inst Oswaldo Cruz. 2000 Apr;95(2):231–235.

11. Bhaskaran S, Butler JA, Becerra S, Fassio V, Girotti M, Rea SL. Breaking Caenorhabditis elegans the easy way using the Balch homogenizer: an old tool for a new application. Anal Biochem. 2011 Jun 15;413(2):123–132.

12. Wong W, Bai X, Brown A, Fernandez IS, Hanssen E, Condron M, Tan YH, Baum J, Scheres SHW. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. Elife. 2014;3.

13. Gupta RS, Siminovitch L. The molecular basis of emetine resistance in Chinese hamster ovary cells: alteration in the 40S ribosomal subunit. Cell. 1977 Jan;10(1):61–66.

14. Plouffe D, Brinker A, McNamara C, Henson K, Kato N, Kuhen K, Nagle A, Adrián F, Matzen JT, Anderson P, Nam T-G, Gray NS, Chatterjee A, Janes J, Yan SF, Trager R, Caldwell JS, Schultz PG, Zhou Y, Winzeler EA. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. Proc Natl Acad Sci USA. 2008 Jul 1;105(26):9059–9064.

Figure 1: Development of a luciferase based *in vitro* translation assay in *Plasmodium falciparum.* A. Protocol for preparing lysate for *in vitro* translation assay. B. A plasmid construct to make generate T7 transcripts containing a *P. falciparum* specific 5' and 3' UTR with a firefly luciferase open reading frame. Maxipreps of the plasmid were digested with PvuII and BamHI to create the T7 transcripts containing only the specific UTRs and luciferase mRNAs. C. Lysates were incubated in the presence of a 10x translation buffer and T7 luciferase mRNAs for a time course of 30 minutes to 120 minutes followed by the addition of luciferin reagent to assay for luciferase activity. The last column of the panel is the same conditions as the 120' lysate but 3xIC50 of cycloheximide was added prior to the start of incubation.

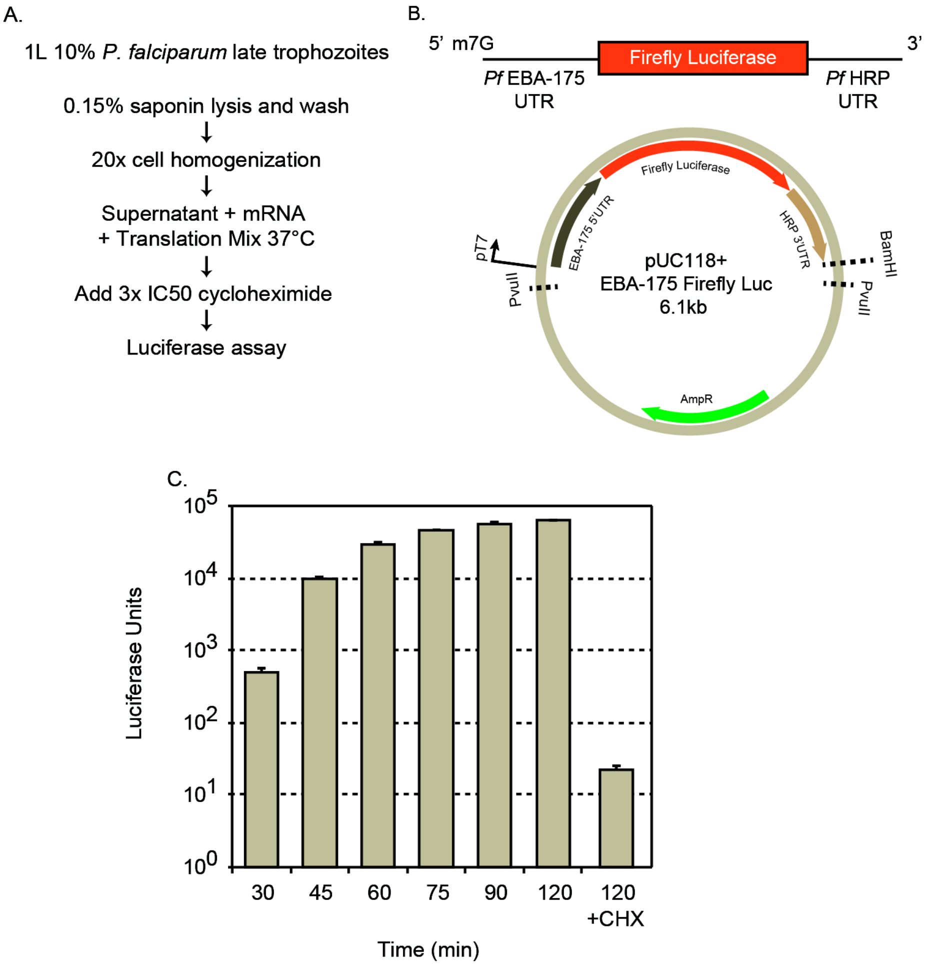

A.
1L 10% *P. falciparum* late trophozoites
0.15% saponin lysis and wash
20x cell homogenization
Supernatant + mRNA
+ Translation Mix 37°C
Add 3x IC50 cycloheximide
Luciferase assay
B.
5' m7G
Firefly Luciferase
3'
Pf EBA-175
UTR
Pf HRP
UTR
Firefly Luciferase
EBA-175 5'UTR
HRP 3'UTR
pT7
PvuII
BamHI
PvuII
pUC118+
EBA-175 Firefly Luc
6.1kb
AmpR
C.
Luciferase Units
10^5
10^4
10^3
10^2
10^1
10^0
30
45
60
75
90
120
120
+CHX
Time (min)

Figure 2: *in vitro* translation assay drug screens. A. Lysates were incubated in the presence of antimalarials and cycloheximide, a general eukaryotic translation inhibitor. Error bars are the standard deviation among 3 biological replicates. B. The average of 3 biological replicates were used to determine the extent of translation inhibition and normalized to the average of the no drug controls present in each plate. Each point on the graph is the response of a single drug. The histogram on the right of the graph displays the total percentage of compounds with the given response.

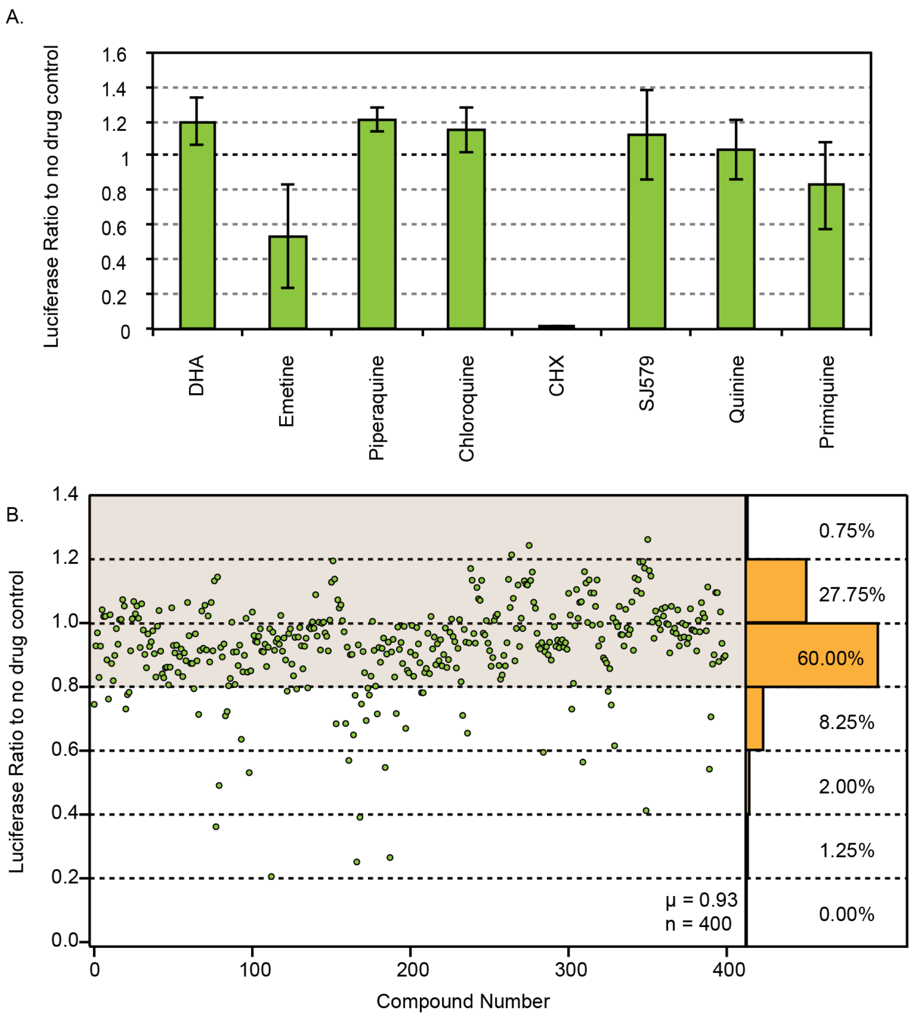
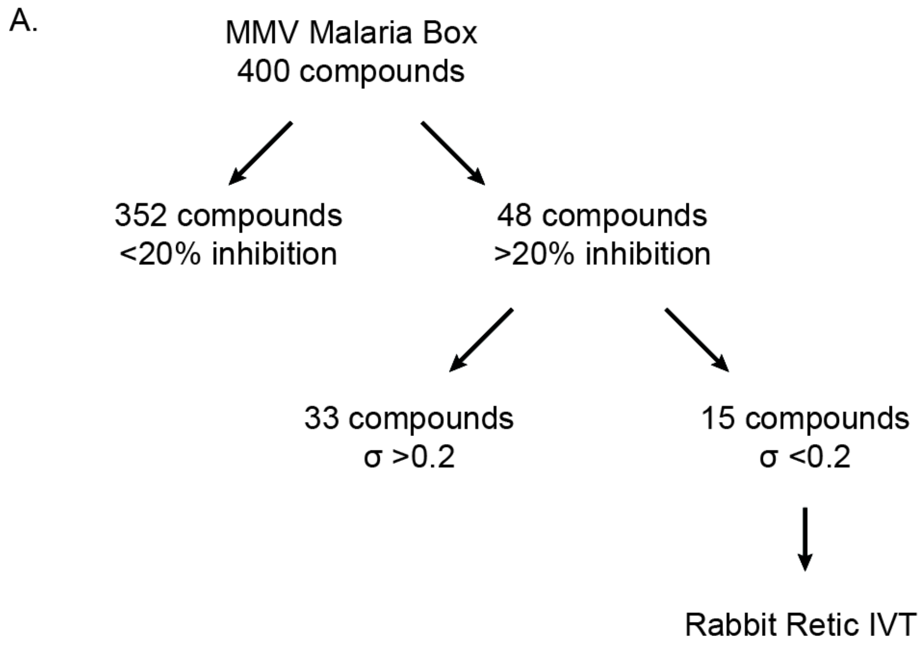


Figure 3: Flow diagram and results of the Malaria Box screen. A. Starting with 400 compounds, each compound was tested in three independent biological replicate *in vitro* translation assays with 1 uM of the compound. Of the 48 compounds that showed at least a 20% translation inhibition effect, only 15 had an acceptable standard deviation of <20%. These 15 were then subject to *in vitro* translation in rabbit reticulocyte lysate, then used to calculate the specificity index of our assay. The 3 compounds with the lowest specificity index will be further characterized for their IC curves. B. The 15 compounds that passed our primary and secondary screens along with their specificity index (*P. falciparum* translation/rabbit reticulocyte translation), reported EC50, and their *P. falciparum* normalized translation in 1uM drug (n=3).



B.

Compound ID	ChEMBL_NTD_ID	CHEMBL EC50 in uM	Specificity Index	Pf average IVT
MMV019124	TCMDC-123658	0.505882901	0.49	0.39
MMV008270	GNF-Pf-780	1.927	0.51	0.27
MMV665886	TCMDC-125438	0.77431091	0.59	0.62
MMV006767	GNF-Pf-3828;TCMDC-123992	0.409;1.129870807	0.62	0.75
MMV007764	GNF-Pf-4338	0.3113	0.63	0.74
MMV666693	TCMDC-124577	0.064397887	0.64	0.71
MMV666057	TCMDC-125853	0.610735136	0.65	0.71
MMV007564	GNF-Pf-4877;TCMDC-124400	0.529;0.747727586	0.66	0.72
MMV019064	TCMDC-123585	0.954986826	0.79	0.80
MMV019110	TCMDC-123639	0.573777396	0.87	0.76
MMV019199	TCMDC-123746	0.933694007	0.87	0.41
MMV007977	GNF-Pf-3717;GNF-Pf-287	1.104;1.853	0.97	0.67
MMV306025	NA	NA	0.99	0.72
MMV666102	TCMDC-123859	0.239309041	1.05	0.21
MMV011522	GNF-Pf-1374;SJ000201287	0.527;15	2.36	0.54

Chapter 6: *cis*-acting Determinants of Translational Efficiency in the Blood stages of *Plasmodium falciparum*

This chapter is a summary of the work done by:

Vida Ahyong and Joseph L. DeRisi

Author contributions:

Vida Ahyong conceived of the project and performed the experiments and analysis. Joseph L. DeRisi supervised the project.

Introduction

Previously we reported a genome-wide characterization of the *Plasmodium falciparum* translation dynamics in five representative stages of the asexual life cycle[1]. This study used the Ribosome Profiling technique to measure transcription and translation by deep sequencing, resulting in a measure of translational efficiency (TE) for every gene in the genome with measureable expression. Our results found that only 10% of the transcriptome is translationally regulated whereas the bulk of genes are transcribed and simultaneously translated in a highly correlated manner. Furthermore, we observed that the 10% of genes that exhibit translational regulation are enriched for functions associated with merozoite egress. For these minority of genes, what are the molecular determinants of their altered translational efficiency? Work in model eukaryotes and mammalian systems have identified a wide range of possible cis-acting features, mostly represented in the 5' UTR of each mRNA species, that may influence translational efficiency[2],[3].

Our previous studies of *Pf* transcription and translation have established a foundational dataset that define many of these features, including the systematic definition of transcription start and stop sites, annotation of novel and antisense transcription and translation, ribosome occupancy within genes and untranslated regions (UTRs), and upstream open reading frames (uORFs). The 5’ UTRs of *Pf* genes are particularly interesting. They are exceptionally long compared to other eukaryotes, they contain an abundance of start codons, and we find many ribosome footprints within them compared to the 3’ UTR[4]. However, the presence of upstream ribosome footprints does not necessarily imply the existence of an upstream ORF (uORF), nor are they inherently predictive of translation efficiency effects of the downstream ORF. Regardless, uORFs remain of considerable interest. It has been established that short upstream ORFs (uORFs) can function to repress translation or in some select cases enhance translation of the downstream open reading frame (ORF) by acting as a sponge for initiating ribosomes[5][6]. Alternatively they can act as a molecular switch to turn on or off the translation of the ORF during cellular stress or development[7][8][9]. Our analysis of 3’ UTRs showed just a handful of genes where the ribosome was associating with the 3’ UTR and as such we do not expect 3’ UTR association to be a significant determinant of translational efficiency in this organism. We also described a surprising amount of antisense transcription. Whether these antisense transcripts play a role in the translation of the cognate mRNA on the opposite strand remains to be revealed. Our work to date does not suggest widespread translational effects of antisense RNAs, however single cell studies, rather than populations, may unmask such activities. Indeed, other studies have suggested that noncoding and antisense RNAs are widespread regulatory features in *P. falciparum*[10][11][12]. Any one of these features alone can strongly regulate the translation of a gene or many of these features may act in a combinatorial fashion to influence translational efficiency. The goal of this study is to identify and characterize the molecular correlates of *Pf* translational efficiency. By doing so, we aim to build a model of *P. falciparum* specific translation and to validate this model using an *in vitro* translation assay previously developed (See Chapter 5).

Results

Nucleotide composition of the 5' UTR

Our previous work employed a Hidden Markov Model (HMM) to define the 5' and 3' transcriptional ends of every well-expressed gene using the high coverage RNA-seq data we had generated. The results showed long 5' UTRs with the median lengths ranging from 607-815 nucleotides throughout the five stages whereas in other eukaryotes, the 5' UTR is generally short, for example in humans the average length of 5' UTRs has been reported as 210 nucleotides[13]. This result along with the abundance of ribosomes found within the 5' UTR led us to question the role of these features with regard to translational efficiency. As an initial step, we examined the genome-wide nucleotide frequencies surrounding the predicted start codon of each gene. By anchoring all genes to the start codon position as nucleotide position 0, we calculated the nucleotide frequencies upstream and downstream of all start codons (Figure 1A) and find enrichment for uracils starting approximately 70 nts upstream of the start codon. To determine whether this uracil-rich tract contributes to translational efficiency, we examined high and low TE genes (high TE: $\log_2$ TE >1.5 n=127, low TE: $\log_2$ TE <-1.5 n=253) and found that the low TE genes lacked a pronounced enrichment for uracil over the same region, on average (nt positions -70nt to -1: 43.1% uracil average in low TE genes, 52.8% uracil average in high TE genes) (Figure 1B). The Kozak sequence is ostensibly the main *cis*-acting sequence that provides the primary sequence information for the ribosome to initiate translation and in other organisms, it has been characterized as a short sequence of approximately 10 nt surrounding the start codon[14]. To determine whether there are differences in the Kozak sequence for high and low TE genes, we generated WebLogo motifs for these two groups with respect to the region surrounding the predicted start codons [15]. For the low TE genes (Figure 2A) we see the characteristic high AT nucleotide content characteristic of the *P. falciparum* genome. At first glance, the high TE genes motif (Figure 2B) looks similar to the low TE genes, but instead of the alternating adenine and uracil nucleotides, we find that the nucleotides preceding the start codon are enriched for homopolymer of uracils, similar to the result we found in the nucleotide frequencies of the high TE genes. Future studies will require measuring short homopolymer tract sizes (i.e.: 5 or 6nt uracils) across the upstream region to determine whether high TE genes do have enrichment in uracil homopolymers.

Secondary structure predictions

The secondary structures within a given mRNA may have significant effects on the translation efficiency, for example, the secondary structures in yeast mRNAs can stall translation until they are transported to the bud[16]. To determine whether we could detect an association between predicted secondary structure and TE in *Pf*, we used the Centroidfold program which implements several RNA folding algorithms such as CONTRAfold to take in an RNA sequence and output secondary structure prediction and free energy of folding for the given sequence (-ΔG)[17]. We chose two different time points (schizonts and rings) instead of a bulk analysis in TE in case there were stage specific differences in expressed mRNAs and their folding attributes. We analyzed high and low TE genes using several sliding window sizes ranging from 20 nt to

100 nt, and anchored all genes by the start codon position at nt 0 for the A in AUG. For the high TE schizont genes (Figure 3A), we observed a 1.23 kcal/mol maximum decrease in the free energy of folding (-ΔG) directly upstream of the start codon followed by a maximum increase of 0.38 kcal/mol in the -ΔG after the start codon for the largest (100nt) window size. The low TE genes lacked the pronounced landscape surrounding the start codon observed in high TE genes. Furthermore the overall predicted free energy less by 1.28 kcal/mol over the upstream region, and by 1.26 kcal/mol for the coding region at the largest window size. The free energy calculations are influenced significantly by the number of guanine and cytosine positions, and therefore these results may also simply reflect a difference in GC content in both the upstream region as well as the coding region with respect to high and low TE. Notably, the profiles of the high TE genes in schizonts and rings are different from each other, suggesting that the mRNAs from each lifecycle stage may inherent possess different features that influence their translational dynamics.

GC content

To directly assess whether there was a difference of GC content with high and low TE genes, we calculated the GC content 500 nucleotides upstream and downstream of the start codon for the 20% high and low TE genes (Figure 4) at a range of sliding window sizes for both the schizonts and ring time points. Surprisingly we find that the GC content upstream of the start codon differs by less than 1% between the high and low TE genes. However, downstream of the start codon, we see an increase of about 2-3% GC content for the high TE genes compared to the low TE genes suggesting that a modestly higher ORF GC content is associated with higher TE.

Multiple regression model of translational efficiency

Using the correlates described above (nucleotide frequencies, GC content, Kozak sequence) combined with factors measured in our Ribosome Profiling dataset (number of AUGs, UTR length, mRNA expression), we sought to evaluate a multiple regression model as a means to predict TE. A similar study in yeast determined that the main determinant influencing TE was the kozak sequence. [18]. To check for independence, we performed cross correlation analysis on multiple parameters in addition to determining the overall correlation with TE. [Table 2] The final list of factors we tested include: the number of 5' UTR AUGs, the length of the 5' UTR, the GC content of the ORF, the number of consecutive As in 12nt of the Kozak sequence, the total number of As in the 12nt of the Kozak sequence, the expression level of the mRNA (reads per million, rM), the maximum difference in free energy (-ΔG) in the Centroidfold analysis, the largest free energy window, a score based on the position specific scoring matrix (PSSM) of nucleotide frequencies from position -145 to -60, the GC content of -20 to 0, and the GC content, of -50 to 0 for all well expressed genes (rM >32). [Table 2]. For factors that have a high cross correlation with another factor such as the number of 5' UTR AUGs and the 5' UTR length (r = 0.91), we chose only one correlate to include into our model, specifically the one with the greater TE correlate. Our final list of factors we used in our regression analyses included the GC content of the ORF (r = 0.13), expression level of the mRNA (r = 0.12), the PSSM motif of -145 to -60 (r = 0.13), and the GC content from position -50 to 0 (r = -0.13). Given these factors, the best that the multiple regression model would be able to predict if all of these factors were independent and evenly weighted would be the sum total of the Pearson correlations, or r = 0.51, however, because there is cross correlation between the factors (i.e.: GC content of the ORF and mRNA expression r = 0.19), our linear regression model would not achieve this level of performance. Clearly, additional parameters that remain to be discovered must contribute to the overall model of *Pf* translation. An approach that would be complementary to this analysis would include a more gross characterization of UTRs using our *in vitro* translation system. By measuring the luciferase expression controlled by a gene's wild-type UTR and then changing particular parameters, determined from our regression model, we would expect corresponding translational efficiency changes and thus luciferase expression.

A special case of translational efficiency

We discovered a special case of translational efficiency in our Ribosome Profiling dataset that deserved further inspection. Though we found over 36,000 possible uORFs (defined by a stretch of at least 2 amino acids) in the genome, upon further visual inspection, only two genes stood out as containing being a uORF that acts to repress translation of the downstream gene. One, which will not be discussed in this thesis but rather in the thesis of a fellow lab mate, Christine Sheridan, is the gene, *var2csa (PFL0030c),* which contains a uORF which may contribute to repressing translation of a gene implicated in placental malaria. The other gene is a 709 amino acid protein (PFE1550w) containing a SURF1 superfamily domain (herein, this protein will be referred to as Surf1) identified by BLASTP analysis (Figure 5) from amino acids position 407-511 and an E-value of 8.82e-09. This is a protein implicated as a regulator of cytochrome oxidase c and mitochondrial biogenesis and when mutated in humans, causes a fatal encephalopathy of infants called Leigh Syndrome[19][20]. In *Pf*, the 5'UTR of this gene has a 55 amino acid uORF that contains an abundance of ribosome footprint reads throughout the entire uORF sequence (Figure 6). We calculated the translational efficiency of the uORF ($\log_2$ TE uORF = 0.55) and in comparison to the TE of the ORF ($\log_2$ TE ORF = -3.21), this result highly suggests that the uORF represses the translation of the downstream gene. To determine whether the uORF produces a functional protein, we performed a multiple alignment analysis of the 55 amino acid sequence using Geneious to seven *Plasmodium* species[21]. The alignment showed that the amino acid sequence is highly conserved with 64.3% identical sites and 82.2% pairwise identity, suggesting that the uORF protein is functionally conserved throughout *Plasmodium* species (Figure 7). We did not find any homology of the uORF protein sequence with any other protein in the non-redundant (nr) protein sequence database using BlastP or tBlastn after excluding other *Plasmodium spp*., indicating that this sequence is unique and important for the *Plasmodium* genus.

To further assess the contribution of the *cis*-acting sequences on the translation of the downstream gene, we dissected the 5' UTR into several constructs fused to a luciferase reporter and performed an *in vitro* translation assay for each construct. We subdivided the 5' UTR into three distinct portions; the leader sequence that is immediately upstream of the uORF, the uORF itself, and the spacer sequence that is the immediate upstream sequence of the ORF. Four constructs were assayed including the full-length sequence containing the leader, uORF, and spacer, the spacer alone, the leader and uORF fused directly to luciferase, and the leader sequence alone (Figure 8). The assay revealed that the full length, spacer only, and uORF fusion constructs had no translation activity while the leader sequence regained the ability to translate luciferase, indicating that the removal of the uORF and the spacer could relieve the *cis*-acting repression of the SURF1 5'UTR. Additional constructs that will be useful for this analysis include a leader fused to the spacer without a uORF, a full length construct without a start codon for the uORF, a construct where the leader and the spacer are switched, and constructs where the uORF nucleotide sequence is scrambled but the amino acid sequence is retained. Each of these additional constructs will reveal intriguing insights into the sequence specific determinants that control the repression of the SURF1 gene.

SURF1 and uORF Protein expression and antibody generation

Beyond the outstanding questions surrounding the translational efficiency and the mechanisms of repression of the SURF1 protein, we sought to evaluate whether expression of the protein could be detected in culture conditions. If SURF1 protein is an activator of mitochondrial biogenesis, we might expect that it would be derepressed under conditions where mitochondrial function is required. Since asexual cultures are grown in high glucose conditions, these parasites typically only carry a single mitochondrion per parasite, whereas in conditions of low glucose such as in stressed gametocyte cultures, the mitochondria expand and develop into a large clustered organelle around the apicoplast and additionally many of the enzymes of the TCA cycle are upregulated in gametocytes[22]. In order to investigate the cellular expression of both the uORF and SURF1, we sought to generate antibodies against both proteins using peptide antigens. For uSURF1, because the protein is only 55 amino acids, we opted to recombinantly express the protein in *e. coli*. The uSURF1 open reading frame was cloned into a set of Macrolab cloning vectors (courtesy of QB3 Macrolab at UC Berkeley) with either N- or C-terminal 6xHis tags or N- or C-terminal 6xHis-MBP tags. The only successful expression was achieved using the N-terminal 6xHis-MBP tagged uORF after a 2h IPTG induction. Protein was purified by batch binding to Qiagen Ni-NTA Agarose beads (Qiagen, Redwood City, CA). The bound protein was successfully eluted from the beads and the 6xHis-MBP tag was then removed using AcTev Protease (Life Technologies, Grand Island, NY) incubation for 3 hours (Figure 9). However, our yield post-cleavage was very low of the purified uSURF1 protein and we struggled to keep the protein soluble as concentrating or leaving the protein in the cold resulted in precipitation. Current efforts to optimize the purification of the uORF protein are underway using different tagged constructs to increase yield. The antibody for the SURF1 protein was made by Pacific Immunology (Ramona, CA) using the peptide sequence Cys-RNNLYDNIKRKEKEEYKNSIE, which is located on amino acid sequence positions 359-379. Efforts are currently underway to validate that the peptide antibody is correctly recognizing the SURF1 protein *in vivo*.

Discussion

Cis-acting determinants of translational efficiency

Our attempt to describe the *cis*-regulatory sequences that correlate with translational efficiency measures resulted in a number of predictors that appear to be associated with translational efficiency. These measures rely on primary and secondary sequence features to explain how a transcript is detected by an initiating or elongating ribosome. However our analysis at best describes only a portion of the story as the best Pearson correlation coefficient we achieved was $r = 0.51$. What other factors could we include to improve this model? One suggestion would be to use an alternative secondary structure folding program such as TEISER that differs from Centroidfold by applying *in vivo* and experimentally determined information of structural elements instead of relying solely on free energy of folding measurements[23]. Another suggestion is to incorporate features of the 3' UTR that may also play a role in translational efficiency that we have yet to fully appreciate such as the length of the poly A tail. Additionally, our model does not take into account RNA binding proteins that act in *trans* by preventing the ribosome from associating with the mRNA, thus searches for RNA binding protein motifs within the primary transcript sequence could improve the model as well. Despite our efforts to discover more predictors, our analysis may be affected by changes in bulk translation or RNA modifications that we cannot detect given the type of information we have. Future studies using our *in vitro* translation assay to directly assess the contribution of each *cis*-acting factor independently will prove to be instrumental in improving this model.

The regulation of SURF1

It is curious why strong uORF repression is only readily apparent in two genes in all the ~5,000 genes in *P. falciparum* despite there being thousands of possible uORFs throughout all 5' UTR sequences. Our current model is that this uORF must act as a fast molecular switch to turn off repression when mitochondrial biogenesis is required such as in conditions where glucose is low and ATP cannot be generated through glycolysis, as is the case during malarial fevers and in the mosquito vector. A recent study that supports this hypothesis found that in *P. berghei,* the mitochondrial ATP synthase is dispensable in asexual blood stages but is required for the sexual stages in mosquito[24]. If this is the case, then why is the uORF amino acid sequence so well conserved among *Plasmodium* species? Does the uORF protein functional beyond its role as an upstream translational repressor? It could be the case that the function of the uORF protein is analogous to puromycin, an antibiotic that causes premature chain termination of the newly formed peptide chain[25]. Validated antibodies to both proteins will allow evaluation of their respective expression dynamics in a range of conditions, including low and high glucose, gametocytogenesis, and during atovoquone treatment which inhibits the mitochondrial electron transport chain[26]. By understanding the mechanisms behind the uORF regulation of this important mitochondrial protein, we will gain insights to the control of mitochondrial biogenesis in *P. falciparum*. Our future studies will focus on how to disrupt this molecular switch to determine what the consequences are for the parasite in the hopes of developing a novel antimalarial therapeutic.

Materials and Methods

Generating nucleotide frequencies and GC content measures

Using the PlasmoDB version 9.1 release of the genome, we retrieved all ORF nucleotide sequences and the 500 nucleotides upstream of the start codons. Using custom python scripts, we aligned all sequences with the nucleotide position 0 as the A in the AUG start codon and calculated the nucleotide frequencies for the entire dataset. GC content measures were performed similarly. We further subdivided the dataset by TE and re-generated the graphs.

Position specific scoring matrix

Using the nucleotide frequency tables we made a position weight matrix (PWM) for the highest TE genes ($\log_{po}$ TE >1.5) using this formula:

$$M_{k,j} = \frac{1}{N}\sum_{i=1}^{N} I(X_{i,j} = k),$$

where N is the number of aligned sequences, Xi,j is the frequency of a nucleotide i at position j, and the M is the matrix with the sum of all probabilities. Then we take the log likelihood ratio LLR = $\log_2$((sum of the weigh matrix/sum of the background frequencies)) of the sum of nucleotide scores across the PWM divided by the background nucleotide frequency (where the background frequency is 0.05 for Gs and Cs and 0.45 for As and Ts in the case of intergenic sequences being 90% AT rich) as a measure of how well a sliding window sequence matches the high TE matrix.

Secondary structure predictions using Centroidfold

Using a local install of Centroidfold, we ran the CONTRAfold program on the 1000 nts surrounding the start codon of every gene with the following parameters: a 5 nucleotide offset, weight of 4 base-pairs, with varying sliding windows. All free energy scores were averaged per window bin for each set of genes analyzed and graphed against their centered nucleotide position.

P. falciparum culturing and in vitro translation

Strains of W2 were maintained in Hyperflasks (Corning, Corning, NY) in 500 mL RPMIc (RPMI 1640 media supplemented with 0.25% Albumax II (GIBCO, Grand Island, NY), 2g/L sodium bicarbonate, 25mM HEPES (pH 7.4), 0.1 mM hypoxanthine, and 50 ug/L gentamycin) in a 37°C, 5% O2, 5% CO2 incubator in 2% hematocrit (HC). Cells were synchronized with 5% sorbitol treatment for 2 generations to achieve high synchronicity.

Parasite cultures were harvested at the late trophozoite stage to approximately 15% parasitemia by centrifugation for 5 min at 1500xg at room temperature and 0.1% final saponin in Buffer A (20 mM HEPES pH 8.0, 2 mM Mg(OAc)2, 120 mM KOAc). Saponin lysed pellets were centrifuged at 4°C 10,000xg for 10 min and washed once with ice cold Buffer A. The pellet was resuspended in 2 mL of Buffer B2 (20 mM HEPES pH 8.0, 100 mM KOAc, 0.75 mM Mg(OAC)2, 2mM DTT, 20% glycerol), flash frozen, and stored in -80°C freezer until the sample was ready to homogenize.

Frozen pellets were thawed on ice and added to a 3 mL luer lock syringe, locked onto a pre-chilled cell homogenizer (Isobiotec, Germany) on ice and passed between two syringes 20 times. Lysate was centrifuged at 4°C 16,000xg for 10 minutes and the supernatant was saved at -80°C until ready for the assays.

In vitro translation reaction were carried out with the following components in 20 uL: 16 uL lysate, 1 ug T7 transcribed firefly luciferase mRNA, 10 μM amino acid mixture, 20 mM HEPES/KoH pH 8.0, 75 mM KoAc, 2 mM Mg(OAc)2, 2 mM DTT, 0.5 mM ATP, 0.1 mM GTP, 20 mM creatine phosphate, 0.2 ug/ul creatine kinase for 1.5h at 37°C. After incubation, the reactions were quenched with 0.5 μM cycloheximide. Reactions were assayed using the Promega GloMax.

Surf1 luciferase constructs and protein expression

Constructs used for T7 transcription and *in vitro* translation used the pUC118 firefly luciferase construct (See Chapter 5 Materials and Methods) as the backbone, replacing the EBA-175 5' UTR with the specified inserts by In-Fusion HD cloning the vector backbone with the PCR amplified and purified 5'UTR (oligos shown in Table 1), incubated for 15 min at 50°C with 0.5uL In-Fusion HD Enzyme mix and a 1:4 ratio of vector to insert, 2 min on ice, followed by transformation into Stellar competent cells. All constructs were Sanger sequence verified.

Constructs for uSURF1 protein expression used a double stranded gene block (uORF_LIC_v2 oligo shown in Table 1) synthesized by IDT (San Diego, CA) to insert into the Macrolab cloning vectors 1B (His6-tev), 1C (His6-MBP-N10-tev), 1G (His6-GST-tev), 2Bc-T (yORF-tev-His6), and 2Cc-T(yORF-tev-MBP-His6). Initial plasmids were cut using Ssp1 for 2 h at 37°C and precipitated. 200ng of the cut vector and 13.3 ng of the dsDNA insert were added to 0.5uL of In-Fusion HD, incubated at 50°C for 15 min followed by 2 min on ice and transformation into Stellar competent cells. All constructs were Sanger sequence verified.

Figure 1: Nucleotide frequencies of the 5' UTR in the *P. falciparum* genome.

All genes were aligned to their start codon position (the A in AUG being in the 0 position) and the nucleotide frequencies were calculated by position looking 200 nucleotides upstream and downstream. A. Nucleotide frequencies for all genes. B. Nucleotide frequencies for either high TE genes or low TE genes.

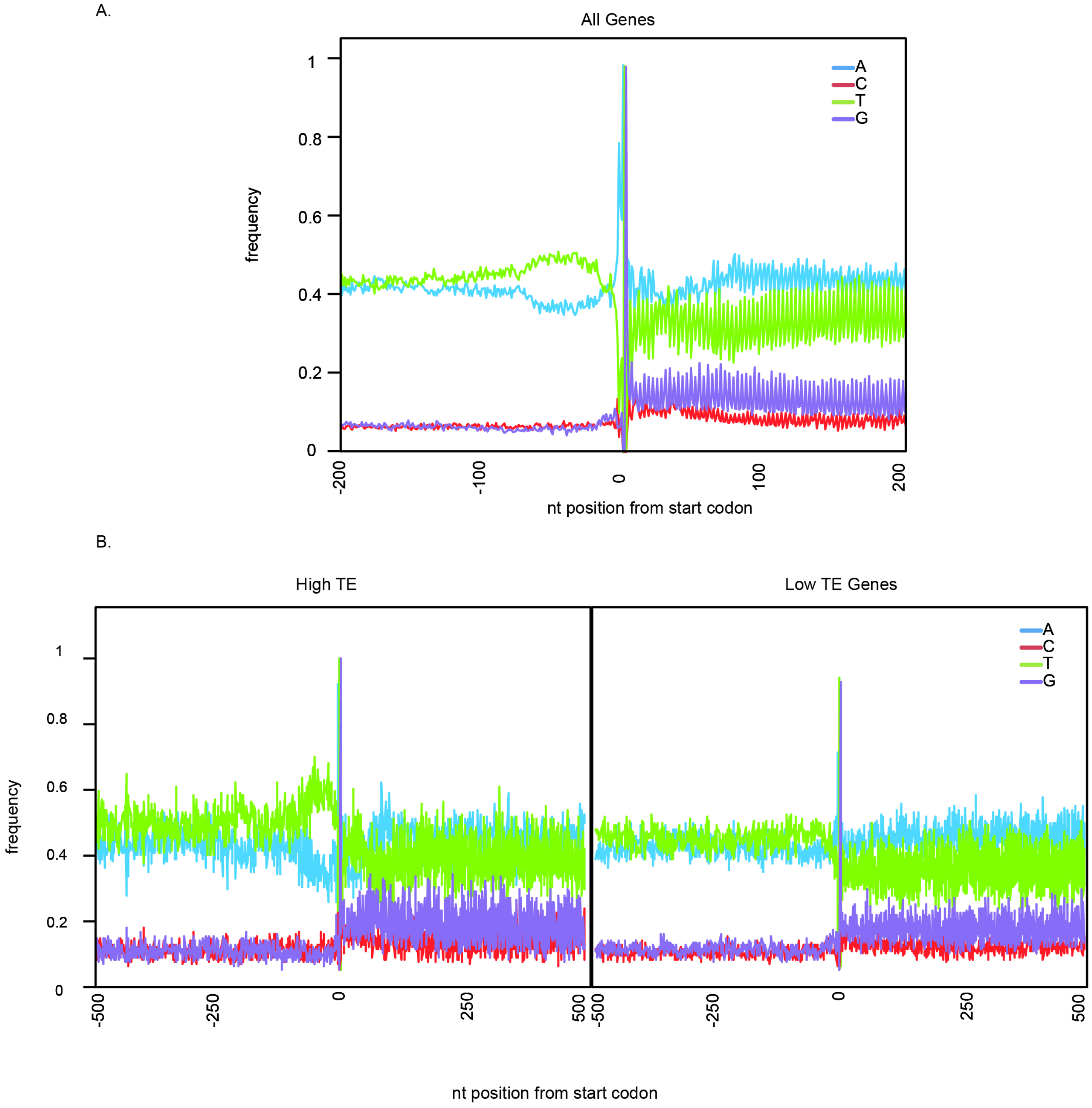
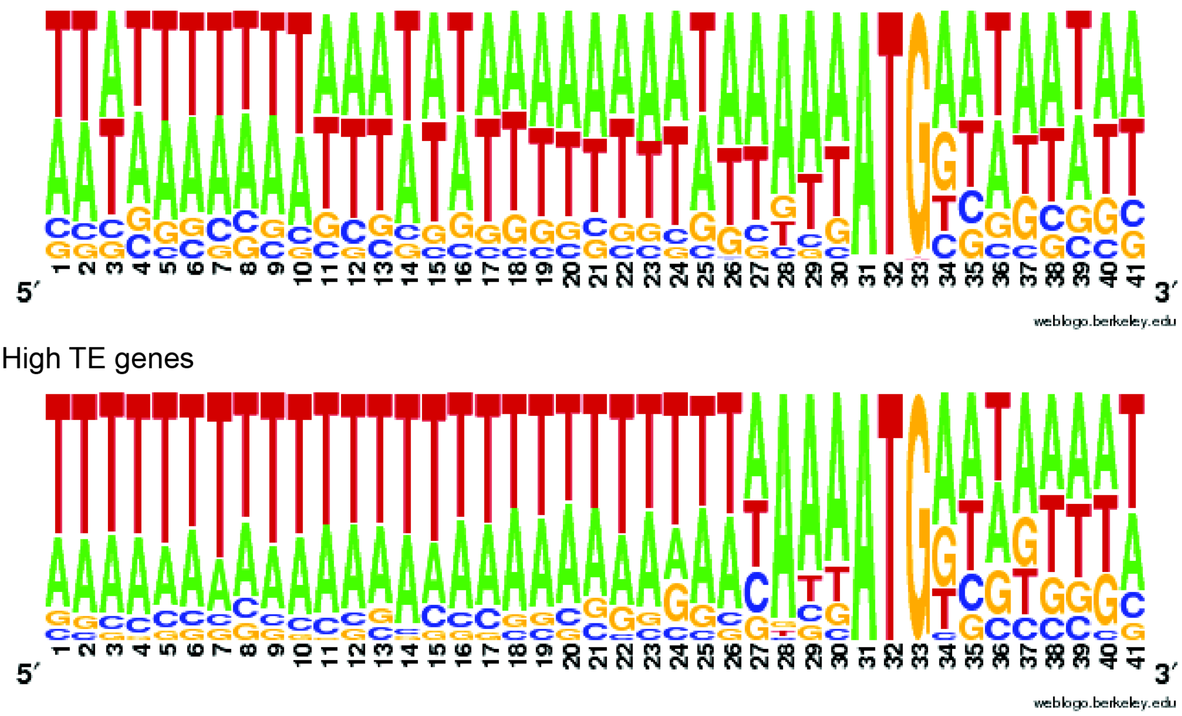


Figure 2: Weblogos of the kozak motifs for high and low TE genes.

A. Low TE genes



B. High TE genes

Figure 3: Free energy measurements with different sliding windows for schizont and ring stage, high and low TE genes. The 20% high and 20% low TE genes were used to run centroidfold to measure the free energy of folding with different size sliding windows. A. schizont stage analysis. B. ring stage analysis.

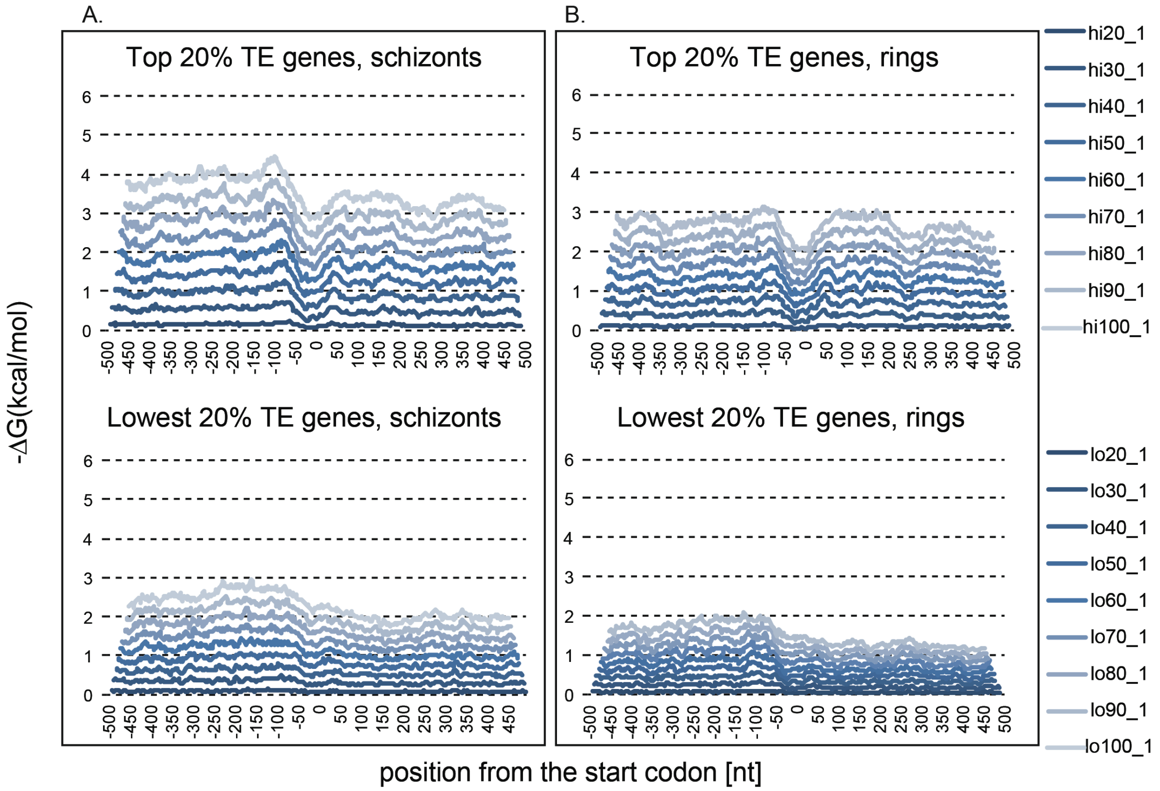
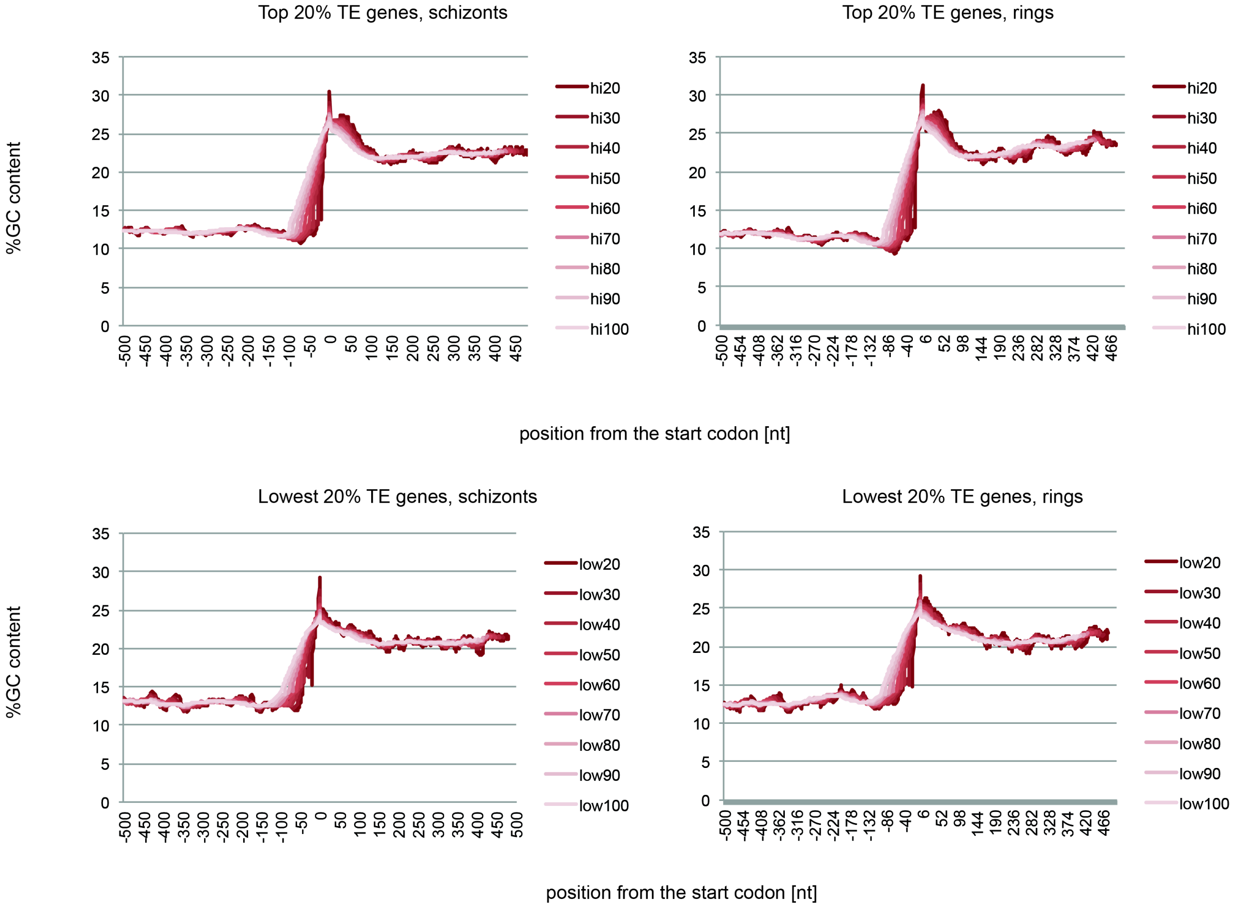


Figure 4: GC content measurements. GC content with multiple sliding windows of high and low TE genes from the schizont and ring stage.



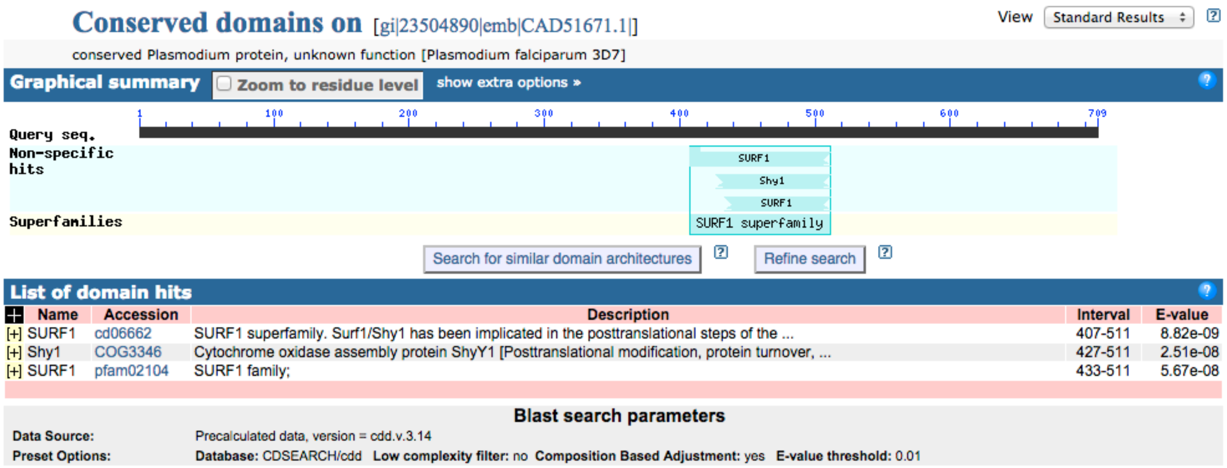
Top 20% TE genes, schizonts

Top 20% TE genes, rings

Lowest 20% TE genes, schizonts

Lowest 20% TE genes, rings

Figure 5: BLASTp results of the amino acid sequence from PFE1550w.



Conserved domains on [gi|23504890|emb|CAD51671.1|]

conserved Plasmodium protein, unknown function [Plasmodium falciparum 3D7]

List of domain hits

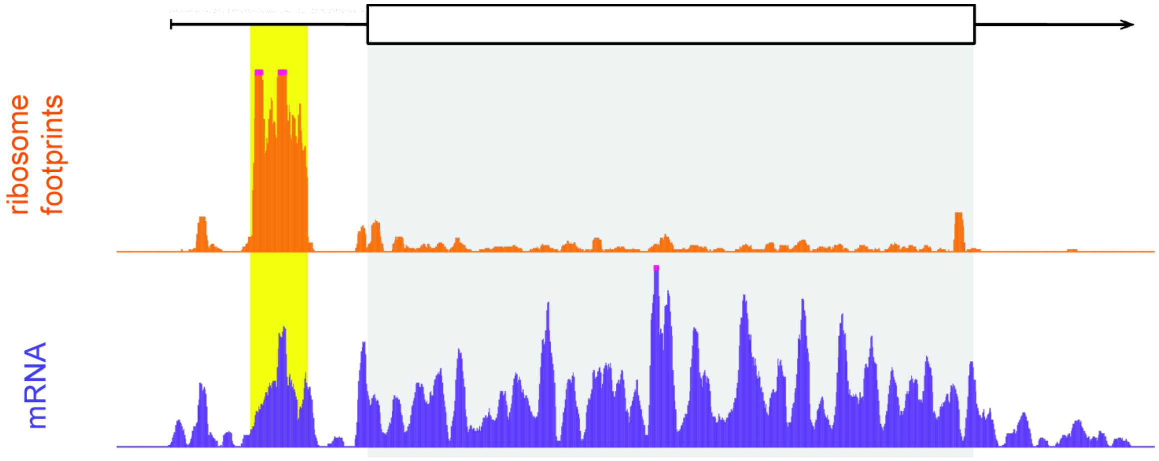
Name	Accession	Description	Interval	E-value
[+] SURF1	cd06662	SURF1 superfamily. Surf1/Shy1 has been implicated in the posttranslational steps of the ...	407-511	8.82e-09
[+] Shy1	COG3346	Cytochrome oxidase assembly protein ShyY1 [Posttranslational modification, protein turnover, ...	427-511	2.51e-08
[+] SURF1	pfam02104	SURF1 family;	433-511	5.67e-08

Blast search parameters

Data Source: Precalculated data, version = cdd.v.3.14

Preset Options: **Database:** CDSEARCH/cdd **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01

Figure 6: Read coverage over PFE1550w (SURF1). Ribosome footprint and mRNA read coverage for the SURF1 locus contain a 55 amino acid uORF (yellow) upstream of the ORF (grey). The TE measurements of the uORF and ORF are shown.



log$_2$TE uORF = 0.55   log$_2$TE ORF = -3.21

Figure 7: Multiple alignments of *Plasmodium spp*. uORF of PFE1550w (SURF1).

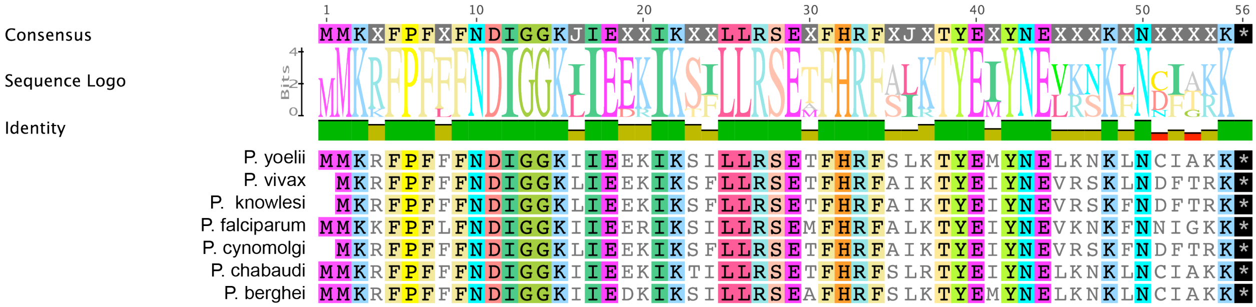

1
10
20
30
40
50
56
Consensus
MMKXFPFXFNDIGGKJIEXXIKXXLLRSEXFHRFXJXTYEXYNEXXXKXNXXXXK*
Sequence Logo
Bits
0
2
4
Identity
P. yoelii
MMKRFPFFFNDIGGKIIEEKIKSILLRSETFHRFSLKTYEMYNELKNKLNCIAKK*
P. vivax
MKRFPFFFNDIGGKLIEEKIKSFLLRSETFHRFAIKTYEIYNEVRSKLNDFTRK*
P. knowlesi
MKRFPFFFNDIGGKLIEEKIKSFLLRSETFHRFAIKTYEIYNEVRSKFNDFTRK*
P. falciparum
MMKKFPFLFNDIGGKIIEERIKSILLRSEMFHRFALKTYEIYNEVKNKFNNIGKK*
P. cynomolgi
MKRFPFFFNDIGGKLIEEKIKSFLLRSETFHRFAIKTYEIYNEVRSKFNDFTRK*
P. chabaudi
MMKRFPFFFNDIGGKIIEEKIKTILLRSETFHRFSLRTYEIYNELKNKLNCIAKK*
P. berghei
MMKRFPFFFNDIGGKIIEDKIKSILLRSEAFHRFSLKTYEMYNELKNKLNCIAKK*

Figure 8: Constructs of the 5'UTR of SURF1 to assay the *cis*-regulatory sequences influencing translational efficiency. Constructs of the leader alone, the spacer alone, the uORF alone, and the full length sequence, all fused the firefly luciferase were assayed for the production of luciferase. The bottom panel shows the relative luciferase units normalized to a control 5'UTR of EBA-175 construct which does not show translational repression of luciferase.

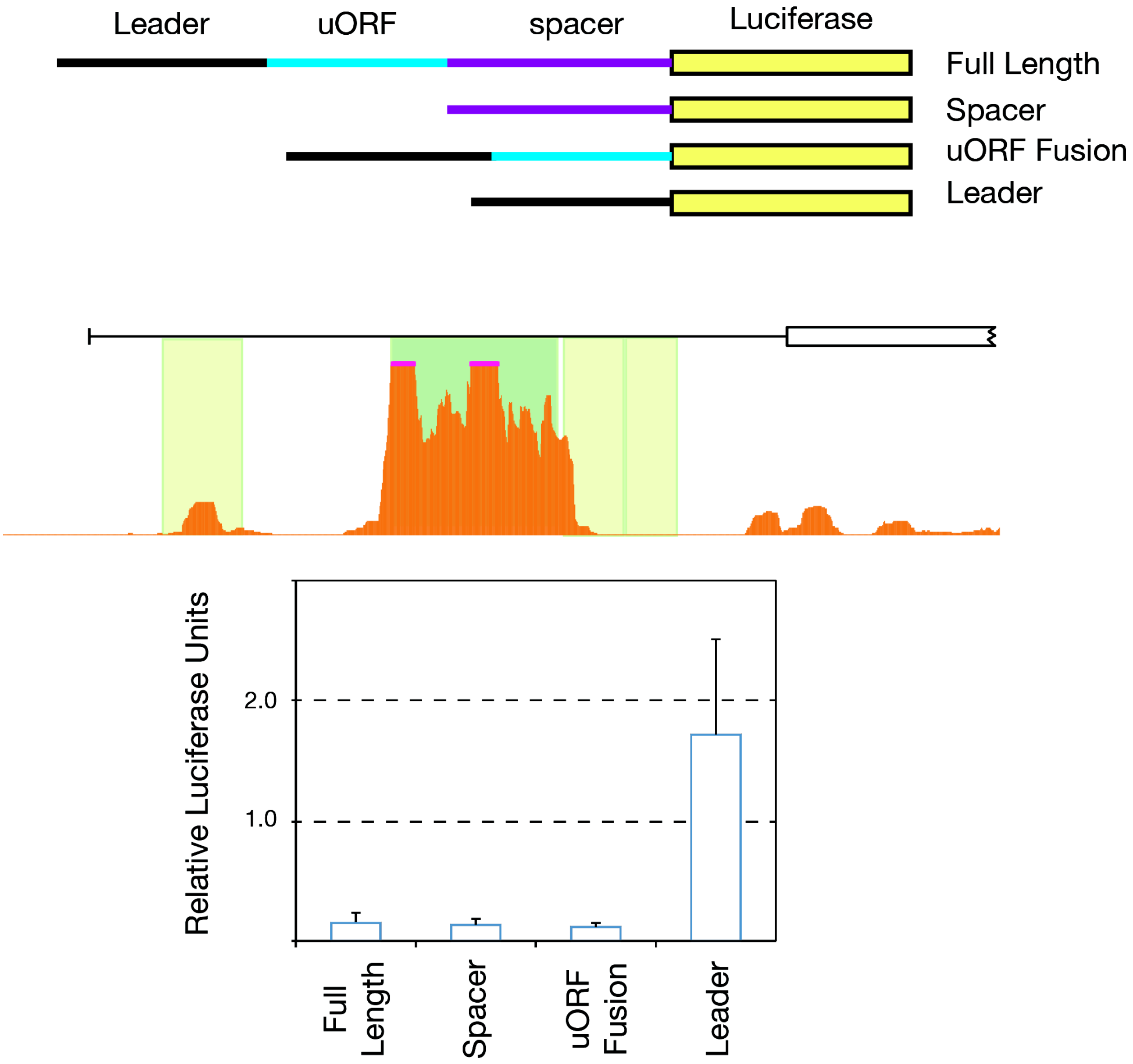


Figure 9: Expression of the uORF of SURF1 in *e. coli*. The *e. coli* containing the uORF construct was induced with IPTG for 2 hours before harvesting. Shown below are the results of the Ni-NTA purification of the 6xHis-MBP-uORF construct. Following TEV cleavage for 3 hours, the 6xHis-MBP is completely cleaved from the uORF protein, resulting in a ~7aa protein.

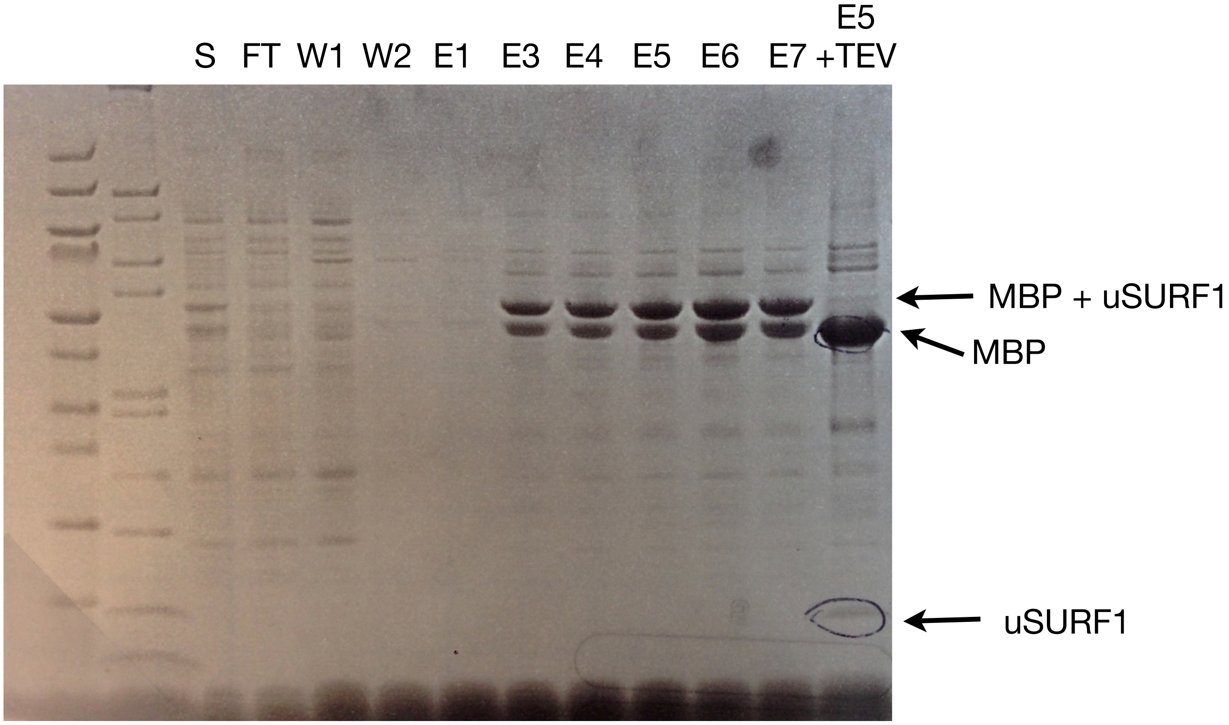
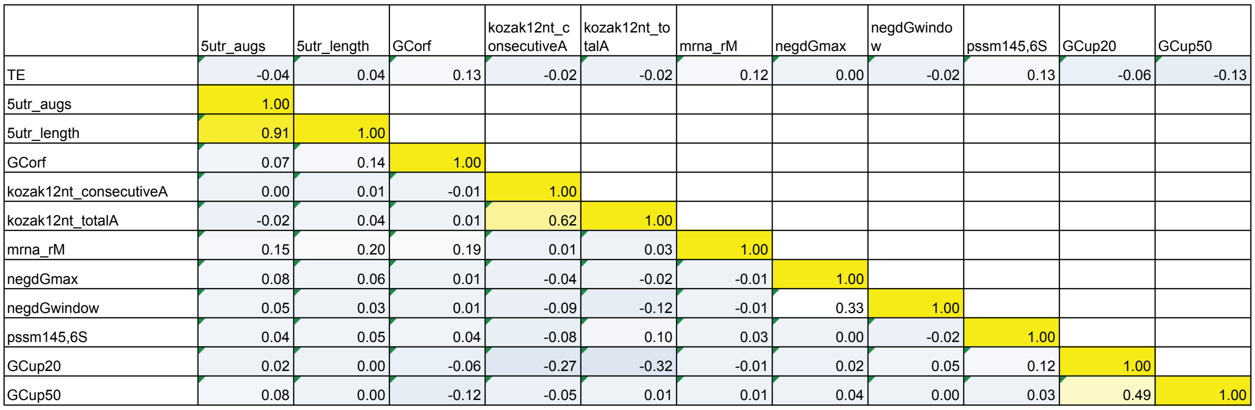


Table 1: Oligos used in this study

oligo name	Short description	Long Description	sequence 5' ->3'
oVA319	F2_F_pfe1550w	to pcr up the upstream sequence of pfe1550w	GCGCAAAATTACTTTGAATACATTAAGTTG
oVA320	RC_pfe1550wORF	to pcr up the upstream sequence of pfe1550w	TGCCTTAAGATCATACTTAAATAAGG
oVA317	Firefly_Fusion_RC	to PCR up the end of the uORF for a direct fusion to firefly	GTTTTTGGCGTCTTCCATTTTTTTCCCTATATTATTAAATTTATTCTTCAC
oVA318	Renilla_Fusion_RC	to PCR up the end of the uORF for a direct fusion to renilla	CATAAACTTTCGAAGTCATTTTTTTTCCCTATATTATTAAATTTATTCTTCAC
oVA332	uUorf_renilla_f	upstream of uorf pfe1550w_RLUC	CGAATTCTAATACGACTCACTATAGGGTTGATTTTTTATAATTATATTATATATATATATA TATATATATATTAATTTAATTATTTATTTTTTTGTTTATATATACAAATATATATTTTTTTTTA TGACTTCGAAAGTTTATGATCC
oVA333	uUorf_renilla_rc	upstream of uorf pfe1550w_RLUC	GGATCATAAACTTTCGAAGTCATAAAAAAAAAATATATATTTGTATATATAAACAAAAAA ATAAATAATTAAATTAATATATATATATATATATATATATATAATATAATTATAAAAAAATCAACC CTATAGTGAGTCGTATTAGAATTCG
oVA314	T7_SpacerF	to PCR up the spacer with T7 homology	TAATACGACTCACTATAGGGGCTATTTCAATAATAAGGGAGG
uORF_LIC_ v2	full uORF for LIC cloning	use this oligo to insert into the LIC vectors	TTTAAGAAGGAGATATAGATCATGATGAAAAAGTTTCCATTTTTATTTAACGATATTG GAGGAAAAATAATTGAAGAAAGAATAAAGTCCATTTTATTAAGAAGTGAAATGTTTC ATAGATTTGCCTTAAAAACTTATGAGATATATAATGAAGTGAAGAATAAATTTAATAAT ATAGGGAAAAAATAATAAGATCCCAACTCCATAA

Table 2: Cross correlations and TE correlations for factors that may affect translational efficiency.



	5utr_augs	5utr_length	GCorf	kozak12nt_consecutiveA	kozak12nt_totalA	mrna_rM	negdGmax	negdGwindow	pssm145,6S	GCup20	GCup50
TE	-0.04	0.04	0.13	-0.02	-0.02	0.12	0.00	-0.02	0.13	-0.06	-0.13
5utr_augs	1.00										
5utr_length	0.91	1.00									
GCorf	0.07	0.14	1.00								
kozak12nt_consecutiveA	0.00	0.01	-0.01	1.00							
kozak12nt_totalA	-0.02	0.04	0.01	0.62	1.00						
mrna_rM	0.15	0.20	0.19	0.01	0.03	1.00					
negdGmax	0.08	0.06	0.01	-0.04	-0.02	-0.01	1.00				
negdGwindow	0.05	0.03	0.01	-0.09	-0.12	-0.01	0.33	1.00			
pssm145,6S	0.04	0.05	0.04	-0.08	0.10	0.03	0.00	-0.02	1.00		
GCup20	0.02	0.00	-0.06	-0.27	-0.32	-0.01	0.02	0.05	0.12	1.00	
GCup50	0.08	0.00	-0.12	-0.05	0.01	0.01	0.04	0.00	0.03	0.49	1.00

Figure legend:

5utr_augs: # of AUGs in the 5'UTR

5utr_length: length of the 5' UTR

GCorf: % GC content of the open reading frame

Kozak12nt_consecutiveA: # of consecutive A nts in the 12nt of the kozak sequence

Kozak12nt_totalA: # of total A nts in the 12nt kozak sequence

mRNA_rM: the expression level (reads per million) from the ribosome profiling RNA-seq

negdGmax: maximum free energy measurement in the 500 nt upstream of the start codon

negdGwindow: maximum free energy measurement window in the 500nt upstream of the start codon

pssm145,6s: position specific scoring matrix consensus score for -145 to +6nt

GCup20: % GC content 20nt upstream of the start codon

GCup50: %GC content 50nt upstream of the start codon

References

1. Caro F, Ahyong V, Betegon M, DeRisi JL. Genome-wide regulatory dynamics of translation in the Plasmodium falciparum asexual blood stages. Elife. 2014;3.

2. Rubio CA, Weisburd B, Holderfield M, Arias C, Fang E, DeRisi JL, Fanidi A. Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. Genome Biol. 2014;15(10):476.

3. Barbosa C, Peixeiro I, Romão L. Gene expression regulation by upstream open reading frames and human disease. PLoS Genet. 2013;9(8):e1003529.

4. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001 Oct 3;276(1-2):73–81.

5. Meijer HA, Thomas AAM. Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. Biochem J. 2002 Oct 1;367(Pt 1):1–11.

6. Wethmar K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdiscip Rev RNA. 2014 Dec;5(6):765–778.

7. Spriggs KA, Bushell M, Willis AE. Translational regulation of gene expression during conditions of cell stress. Mol Cell. 2010 Oct 22;40(2):228–237.

8. Bancells C, Deitsch KW. A molecular switch in the efficiency of translation reinitiation controls expression of var2csa, a gene implicated in pregnancy-associated malaria. Mol Microbiol. 2013 Nov;90(3):472–488.

9. Yaman I, Fernandez J, Liu H, Caprara M, Komar AA, Koromilas AE, Zhou L, Snider MD, Scheuner D, Kaufman RJ, Hatzoglou M. The zipper model of translational control: a small upstream ORF is the switch that controls structural remodeling of an mRNA leader. Cell. 2003 May 16;113(4):519–531.

10. Vembar SS, Scherf A, Siegel TN. Noncoding RNAs as emerging regulators of Plasmodium falciparum virulence gene expression. Curr Opin Microbiol. 2014 Aug;20:153–161.

11. Broadbent KM, Broadbent JC, Ribacke U, Wirth D, Rinn JL, Sabeti PC. Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. BMC Genomics. 2015;16:454.

12. Amit-Avraham I, Pozner G, Eshar S, Fastman Y, Kolevzon N, Yavin E, Dzikowski R. Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum. Proc Natl Acad Sci USA. 2015 Mar 3;112(9):E982–991.

13. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001 Oct 3;276(1-2):73–81.

14. Kozak M. An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 1987 Oct 26;15(20):8125–8148.

15. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004 Jun;14(6):1188–1190.

16. Hinnebusch AG. Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiol Mol Biol Rev. 2011 Sep;75(3):434–467, first page of table of contents.

17. Sato K, Hamada M, Asai K, Mituyama T. CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W277–280.

18. Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, Koller D. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol Syst Biol. 2014;10:770.

19. Shoubridge EA. Cytochrome c oxidase deficiency. Am J Med Genet. 2001;106(1):46–52.

20. Zhu Z, Yao J, Johns T, Fu K, De Bie I, Macmillan C, Cuthbert AP, Newbold RF, Wang J, Chevrette M, Brown GK, Brown RM, Shoubridge EA. SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome. Nat Genet. 1998 Dec;20(4):337–343.

21. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012 Jun 15;28(12):1647–1649.

22. Okamoto N, Spurck TP, Goodman CD, McFadden GI. Apicoplast and mitochondrion in gametocytogenesis of Plasmodium falciparum. Eukaryotic Cell. 2009 Jan;8(1):128–132.

23. Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, Salavati R, Cristea IM, Tavazoie S. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature. 2012 May 10;485(7397):264–268.

24. Sturm A, Mollard V, Cozijnsen A, Goodman CD, McFadden GI. Mitochondrial ATP synthase is dispensable in blood-stage Plasmodium berghei rodent malaria but essential in the mosquito phase. Proc Natl Acad Sci USA. 2015 Mar 23;

25. Pestka S. Inhibitors of ribosome functions. Annu Rev Microbiol. 1971;25:487–562.

26. Painter HJ, Morrisey JM, Vaidya AB. Mitochondrial electron transport inhibition and viability of intraerythrocytic Plasmodium falciparum. Antimicrob Agents Chemother. 2010 Dec;54(12):5281–5287.

Chapter 7: Conclusions and Future Directions

The advances in genomic technologies have revolutionized the manner that we approach science by permitting unbiased and high-throughput biological measurements. In the DeRisi lab, I heavily relied on deep sequencing approaches to investigate the human malaria parasite, *Plasmodium falciparum,* which requires a unique blend of computational and wet lab expertise and as such, I have devoted a significant amount of my graduate training to developing my proficiency in both. My research has explored topics in antimalarial drug resistance, post-transcriptional gene regulation, drug discovery, and the discovery of *cis*-regulatory features of mRNA.

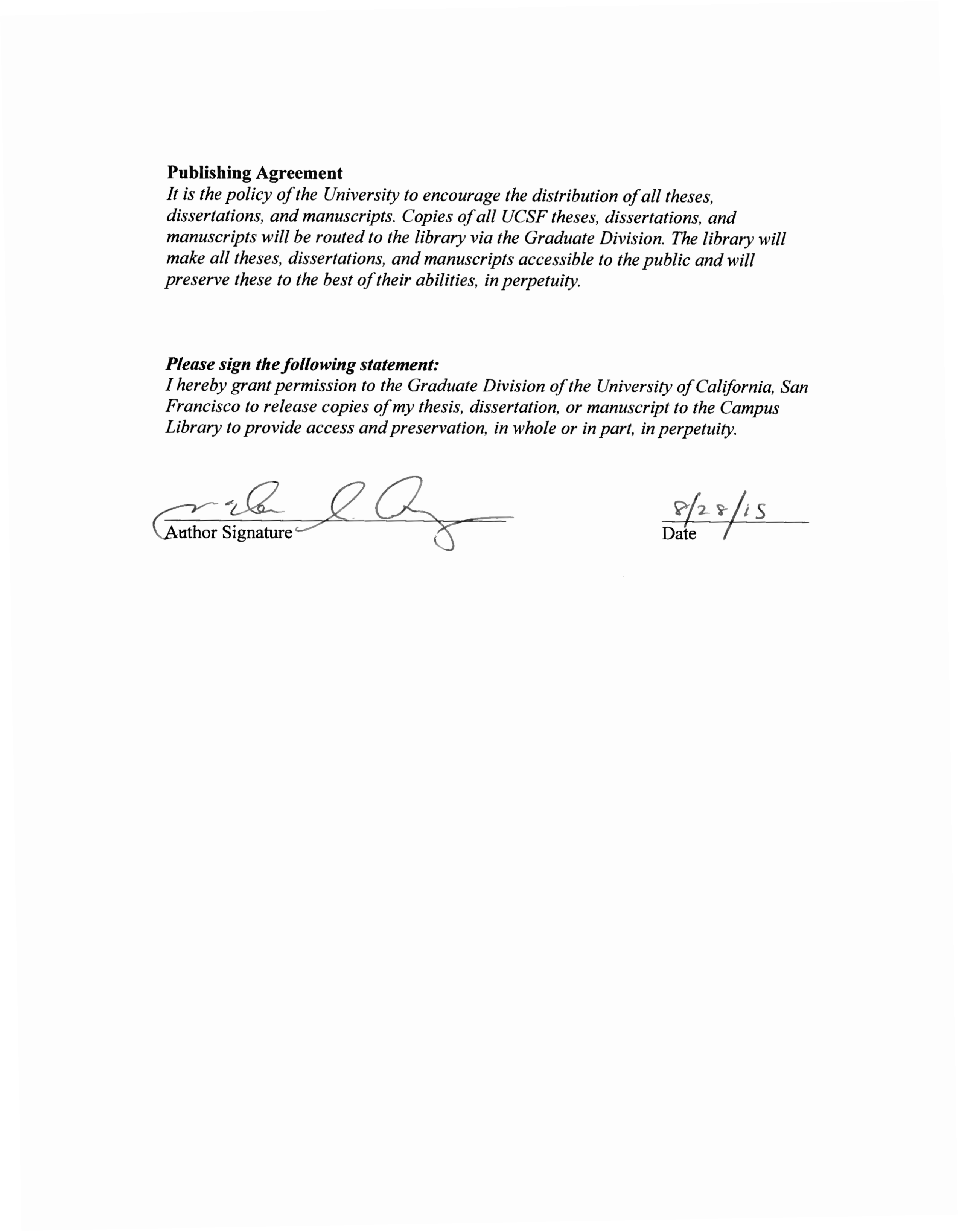
As a young graduate student, I began my training in the development of computational programs to identify genomic mutations that confer resistance to antimalarial drugs. Chapter 2 and 3 demonstrate how valuable deep sequencing technologies are towards attaining this end goal. In Chapter 2, we identified amplifications to a known drug resistance gene that scaled with the level of drug resistance. Additionally we discovered a novel mechanism of genome amplification that was only possible through the analysis of deep sequencing data. We found that the junctions between amplicons are long A/T tracks that are triggers for breakage and homologous recombination-like pathways to expand a portion of the genome. These amplicons are always configured in the same orientation such that de-amplification can occur quickly in cases where drug pressure is relieved, through similar homologous recombination-like pathways. This work describes a novel mechanism that may be a common feature that this parasite uses to escape drug pressure and future work to prevent this mechanism could prove useful in the development of antimalarial combination therapies.

In Chapter 3, I used similar bioinformatics software to find genetic mutations that confer resistance to the very important antimalarial, Artemisinin. Emerging Artemisinin resistance on the Thai-Cambodian border is a threat to global health as Artemisinins are one of the few robust, fast acting antimalarials that are effective worldwide. Though recent studies have implicated the parasite's kelch13 protein and phosphatidylinositol-3-kinase as possible targets of resistance, it is also possible that there are many genes in the target pathway whose mutations can confer drug resistance. The *in vitro* selection of drug resistant parasite in two independent lab strains both point to an amplification of genes on chromosome 10 as conferring resistance to Artemisinin during *in vitro* drug culture. We believe that just one gene within this amplified region is the gene requiring the amplification to escape the drug pressure similarly to the amplification detected in Chapter 2 which covered the DHODH locus. This study and the identification of a narrow set of possible gene targets will be useful in understanding the entire pathway that Artemisinin targets and could influence how new therapies are designed to avoid drug resistance. The DeRisi lab routinely uses deep sequencing to identify more targets of antimalarial drugs and furthermore we have trained a number of researchers in many other labs to use our custom python scripts for similar studies.

For the remainder of my thesis, I focused on the investigation of translation in the asexual blood stages of *P. falciparum*. This study was provoked by suggestions in the field that post-transcriptional regulation was a major theme in the control of gene expression especially in the late schizont stages of the intraerythrocytic developmental cycle. Thus we measured translation on a genome-wide scale using the Ribosome Profiling technique. After performing many replicates to assure ourselves that our data was accurate, we found that throughout the blood stages, transcription and translation are tightly correlated with only ~10% of genes translationally regulated. Though this result was surprising considering the suggestions in the field but nonetheless, we learned that at the end, our data unbiasedly pointed us in the right direction. Additionally, we discovered many new and novel themes in malaria translation. We found that antisense transcription was widespread throughout the blood stages and seem to be produced from most promoters that act bi-directionally. Interestingly, we find ribosome footprints throughout many of these antisense transcripts, though the repercussions will remain to be determined. We also described transcriptional start and stop sites in the most complete manner to date. This information alone will be extremely valuable to researchers who require knowledge of where a transcript begins and ends. Finally, we explored regulation *via* the untranslated regions by quantifying the number of upstream start codons (uAUGs), upstream open reading frames (uORFS), and general ribosome density, all of which could influence the translation of the mRNA. All of this work culminated in a rich resource for the malaria research community and has opened many more questions to be answered.

Future research in our lab will be focused on answering some of the outstanding questions that resulted from our ribosome profiling dataset. We found a special case of translational repression through the presence of a uORF in the SURF1 gene. This gene is important for the regulation of mitochondrial biogenesis and further research into this regulation could prove useful in future drug therapeutics by targeting the ability of the parasite to produce more cellular powerhouses. We developed a high-throughput, luciferase based translation system which allows us to further dissect the regulatory elements from the 5'UTR, in an attempt to understand the *cis*-acting sequences that modulate translation. Furthermore, I began building a multiple regression model that takes in all the quantifiable predictors of translation from our Ribosome Profiling dataset and analysis of the nucleotide sequences of malaria genes with the end goal to produce an *in silico* model malaria translation. Future work will use this model to modulate wild-type transcript sequences to either perform better or worse, depending on the desires of the researcher, when added to an *in vitro* translation assay or transfections with *in vivo* parasite cultures.

In total, the work presented in this thesis contributes to our understanding of the control of gene expression in the medically relevant human malaria parasite. We have developed novel tools to understand how the parasite can mutate their genome to escape death by antimalarial drugs. We also comprehensively characterized the process of translation in the blood stages and provided valuable resources for the research community. And finally, we have discovered novel translation mechanisms of the parasite that can be used as targets for future drug therapeutics.



Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

8/28/15
Date