Transcriptional profiling and network-based gene annotations of human malaria parasite plasmodium falciparum

Hu, Guang An

2009

Hu, G. A. (2009). Transcriptional profiling and network-based gene annotations of human malaria parasite plasmodium falciparum. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/47451

https://doi.org/10.32657/10356/47451

TRANSCRIPTIONAL PROFILING AND NETWORK-BASED GENE ANNOTATIONS OF HUMAN MALARIA PARASITE *Plasmodium falciparum*

GuangAn Hu

School of Biological Sciences

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2009

Declaration

This is to certify that

(1) the thesis comprises only my original work in the School of Biological Sciences of Nanyang Technological University towards the requirement of PhD degree;

(2) due acknowledgement has been made in the text to all other material used;

(3) the thesis does not exceed the word length for this degree.

GuangAn Hu

July 2008

Abstract

Transcriptional profiling and network-based gene annotations of human malaria parasite *Plasmodium falciparum*

By GuangAn Hu

Supervisor: Dr. Zbynek Bozdech

Plasmodium falciparum is the most causative agent of the deadliest form of human malaria responsible for around 2 million deaths in the world. Even 6 years after the genome sequencing, more than 50% genes of *P. falciparum* remain functionally uncharacterized. In this work we generated large functional datasets and systematically analyze these data combining other public functional datasets to characterize gene function, cellular process and gene regulation in *P. falciparum*. We developed the program OligoRankPick which uses a weighted rank-based strategy for the design of long oligonucleotide DNA microarrays. OligoRankPick does not rely on direct oligonucleotide exclusion by parameter cutoffs but instead optimizes all parameters in context of each other. Using this program we have designed several long oligonucleotide DNA microarrays for the parasitic species including *P. falciparum, P. vivax* and pan-rodent malaria. Based on the designed DNA microarray, we report on extensive transcriptional profiling of *P. falciparum* parasites using 20 small molecular compounds including several common antimalarial drugs. Diverse gene responses were observed in different drug or compound treatments and this perturbation data has a high predictive accuracy of functionally related genes based on their transcriptional regulation. Specifically, transcriptional analysis showed specific gene responses induced by inhibiting classic signaling pathways when parasite cells

were treated from the early schizont stage and several transcription factors and signaling genes were up-regulated by the inhibition of calcium dependent signaling and calcineurin signaling pathways, suggesting the phosphorylation and/or dephosphorylation play vital roles of the gene expression regulation in *P. falciaprum*. Combining the transcriptional profile with *in silico* generated phylogenetic profiles, domain-domain interaction evidence and the yeast two-hybrid system-based protein-protein interaction dataset we construct a high confidence gene interactome network using a probabilistic Bayesian network approach. Based on this network, we assign function to 2547 hypothetical proteins using the weighted neighbor counting method (WNC). To demonstrate the utility of this network we assemble a sub-network of genes associated with merozoite invasion and predict 263 new proteins that are associated with this process. The predictions were validated by the localization of a subset of previously uncharacterized protein candidates of the *Plasmodium* invasive form (merozoites) which further confirms their predicted functions in the malaria parasite invasion process.

Table of Contents

List of tables

List of figures

Acknowledgements

First and foremost, I thank Dr. Zbynek Bozdech for his mentorship and for providing an inspiring environment in which to pursue my research interests. I have greatly appreciated his zeal for discovery, elucidation of key questions, and support of conference presentations. I also thank my thesis committee who has also been extremely helpful in contributing their wide-ranging expertise to my interdisciplinary project.

I would like to acknowledge all members of the ZB Lab for their useful suggestions and support. I have thoroughly enjoyed many enthusiastic insights into our research and our discussions on a wide range of topics. I would especially like to thank Yang Ye, Dr. Mu Yuguang and Dr. Lin Xin for the many stimulating discussions and computer programs.

Finally, I am indebted to my wife Zhi Hui for her kindest concern of my life and research.

List of publications

Guangan Hu, Manuel Llinás, Jingguang Li, Peter Rainer Preiser and Zbynek Bozdech. Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. BMC Bioinformatics 2007, 8:350.

Masayo Kotaka, Hong Ye, Reema Alag, **Guangan Hu**, Zbynek Bozdech, Peter Rainer Preiser, Ho Sup Yoon and Julien Lescar. Crystal Structure of the FK506 Binding Domain of *Plasmodium falciparum* FKBP35 in Complex with FK506. Biochemistry, 2008, 47 (22), 5951–5961.

Zbynek Bozdech, Sachel Mok, **Guangan Hu**, Mallika Imwong, Anchalee Jaidee, Bruce Russell, Hagai Ginsburg, Francois Nosten, Nicolas Day, Nicolas White, Jane Carlton, Peter Preiser. The transcriptome of *Plasmodium vivax* reveals evolution of transcriptional regulation in malaria parasites. PNAS, 2008, 105(42):16290-16295.

Guangan Hu, Ana Cabrera, Sachel Shui-li Mok, Maya Kono, Balbir K Chaal, Silvia Haase, Sabna Cheemadan, Brigitta D Wastuwidyaningtyas, Tobias Spielmann, Peter R Preiser, Tim–Wolf Gilberger, Zbynek Bozdech. Transcriptional profiling of growth perturbations and a functional interactome network of human malaria parasites, *Plasmodium falciparum*. Nature biotechnology, revised.

Kinsley Liew, **Guangan Hu**, Zbynek Bozdech, Peter Preiser. Development of a pan-rodent long oligonucleotide chip for the genomic and transcriptomic analysis of the three rodent malalria species *P. berghei*, *P. chabuadi* and *P. yoelii*. manuscript

Meeting reports

Hu Guangan, Bozdech Zbynek. Protein-protein Interactions of P. falciparum Merozoite Invasion Machinery. Molecular Parasitology Meeting XVII, 2007, Woods Hole, USA. (poster)

Liew Kingsley, **Hu Guangan**, Bozdech Zbynek, Preiser Peter. Development of a pan-rodent long oligonucleotide microarray for global genomics and transcriptomic study

of *Plasmodium yoelii, Plasmodium berghei* and *Plasmodium chabaudi*. Molecular Parasitology Meeting XVII, 2007, Woods Hole, USA. (poster)

Guangan Hu, Sachel Shui-li Mok, Balbir Kaur Chaal, Sabna Cheemadan, Brigitta Dewi Wastuwidyaningtyas, Peter R Preiser, Zbynek Bozdech. Network-based gene function predictions in *Plasmodium falciparum*. Singapore Malaria Network Meeting 2008. NTU, Singapore. (talk)

Ana Cabrera, Maya Kono, **Guangan Hu**, Silvia Haase, Tobias Spielmann, Zbynek Bozdech, Tim–Wolf Gilberger. Tagging the invasome: evaluation of a probabilistics sub-network of Plasmodium falciparum genes associated with merozoite invasion. Molecular Parasitology Meeting, 2008, Woods Hole, USA. (poster, "honorable mention" award)

Chapter 1 Introduction

1.1 Human malaria disease

The human malaria parasite is still an affliction on human populations and the incidence of the disease has been increasing in recent years. It is estimated that malaria affects 300 million to 500 million people and kills 1.5–2.5 million people each year, mostly among young children and pregnant women in sub-Saharan Africa (Hay, et al., 2004). Moreover, recent clinical investigations indicated that the original number of the recorded episodes was largely underestimated, especially for outside Africa, and that malaria is spreading with much greater velocity than previously believed (figure 1.1) (Snow, et al., 2005). This situation is caused mainly by resistance to all available chemotherapy and absence of any effective vaccine. Thus development of novel drugs as well as an effective vaccine

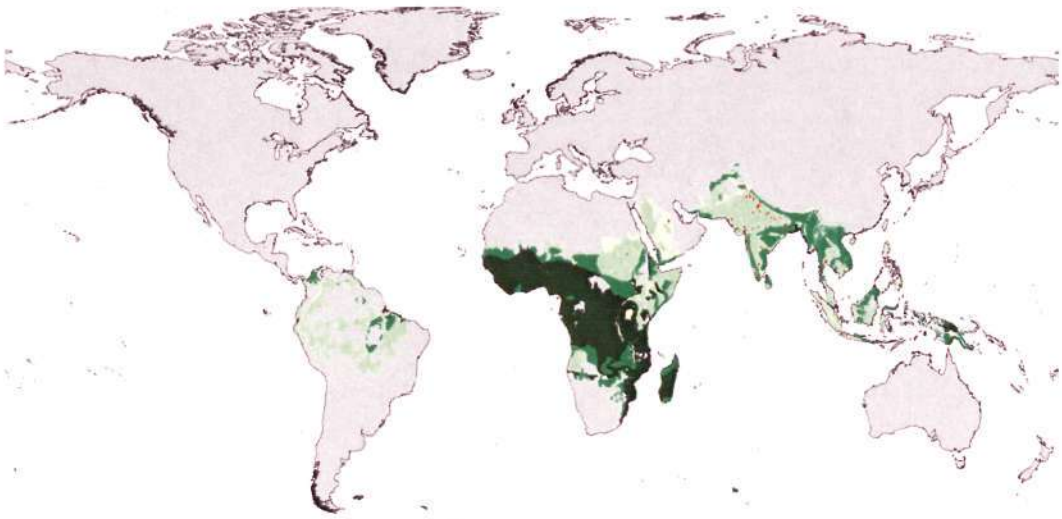


Figure 1.1 The global clinic risk recorded episodes was largely underestimated. The figure was copies from Snow et al. (Snow, et al., 2005).

is crucial for the fight against this deadly disease. To achieve this goal a comprehensive understanding of molecular aspects of human malaria is essential. One of the major tasks in this effort is functional annotation of malaria parasite genes in order to identify and characterize new potential molecular targets for drug and vaccine development.

1.2 *Plasmodium falciparum* life cycle

Human malaria is caused by several species of protozoa parasites, *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae* and *Plasmodium knowlesi*, among which *P. falciparum* is the most lethal form. The *Plasmodium* parasites are characterized by a complex life cycle comprised of a series of dramatic developmental stages taking place in both of its hosts, human and mosquito (figure 1.2). The parasite-infected mosquito bites and injects invasive sporozoites into the human host blood stream. The sporozoites are rapidly sequestered in the liver where after a brief development, and another invasive form of parasite (merozoites) is produced and released into the bloodstream. In the blood, the parasites invade and multiply in the red blood cells. This asexual multiplication is known as intraerythrocytic developmental cycle (IDC) (figure 1.2), which includes three distinct morphological stages: ring, trophozoite, and schizont stage. The IDC is completed in approximately 48 hours, during which the early stages (ring and trophozoite) are highly metabolically active, rapidly ingesting hemoglobin, and performing the majority of generic cellular processes associated

with their growth. This process culminates about 30 hours post invasion (hpi), when the cells start to replicate their DNA (early schizont stage). At approximately 48 hpi, newly formed mature merozoites rupture from the red cell and invade new cells to reinitiate another cycle. During the asexual multiplication, a fraction of parasite cell differentiates into a precursor sexual stage known as gametocytes. The sexual development and fertilization is completed in the mosquitoe gut to reinitiate a new cycle of parasite transmission.

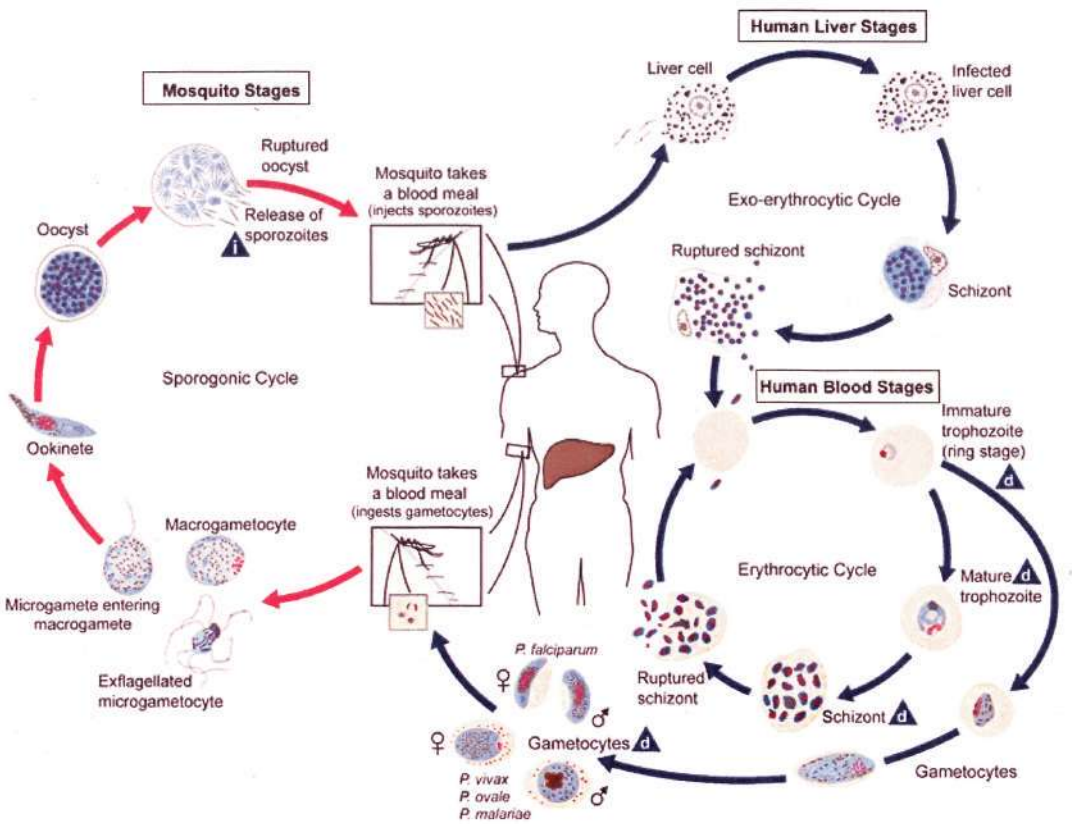


Figure 1.2 The life cycle of *Plasmodium* (The graph was copied from the Parasite Image Library, http://www.dpd.cdc.gov/dpdx/HTML/Image_Library.htm). Sporozoites infect liver cells from female mosquito and mature into schizonts, which rupture and release merozoites (exo-erythrocytic schizogony). Then the merozoites infect red blood cells and begin the asexual blood cycles. Some

parasites differentiate into sexual erythrocytic stages (gametocytes), which are ingested by a mosquito. The parasites' multiplication in the mosquito is known as the sporogonic cycle.

Although all developmental stages are essential for successful parasite transmission and thus progression of the disease in the human population, the asexual blood stage is responsible for all malaria symptoms. The blood stage is also the most important target area for the majority of presently available antimalarial drugs as well as development of new antimalarial strategies. These development efforts were recently enhanced by completion of the genome sequences of several *Plasmodium* species and comprehensive analyses of global transcription profiles during the *Plasmodium* life cycle (Bozdech, et al., 2003; Carlton, et al., 2002; Gardner, et al., 2002; Le Roch, et al., 2003).

1.3 The genome, transcriptome, proteome and interactome of *P. falciparum*

The 22.8 Mb genome of *P. falciparum* is comprised of 14 linear chromosomes ranging in size from 0.64 – 3.3 Mb and two non-nuclear genomes: a circular 35-kb plastid-like genome and a linear 6-kb mitochondrial genome (Waller, et al., 2004). The genome of the 3D7 strain of *P. falciparum* was the first parasite genome to be sequenced to completion (Gardner, et al., 2002). The genome has more than 5300 coding sequences (CDS), however, more than 60% of the predicted genes could

not be assigned functions because they do not have sequence homology with known genes in other organisms (Gardner, et al., 2002). Curiously, only 8% of the *P. falciparum* genes could be assigned functions in metabolism, in contrast to 17% of the genes of the yeast *Saccharomyces cerevisiae* (Kooij, et al., 2006). This suggests that enzymes are more difficult to be identified by sequence homology in *P. falciparum* owing to its great evolutionary distance from other well-studied organisms. Obviously, a lot of enzymes related to metabolic process possibly are present in the 60% hypothetical proteins. Hence development of new methods to characterize these hypothetical proteins is crucial and urgent to understand the biology of malaria.

Comprehensive profiles of transcript levels throughout the complete life cycle of the *P. falciparum* parasite have been extensively investigated. It was shown that each gene is activated specifically at the time when its function is required (Bozdech, et al., 2003; Le Roch, et al., 2003; Llinas, et al., 2006). Although these results brought numerous insights into *Plasmodium* biology, the transcriptome data had only a limited impact on gene anotaion. This is mainly due to the monotonous character of the *Plasmodium* transcriptome where large and functionally diverse groups of genes share common transcriptional profiles across the *Plasmodium* life cycle. This phenomenon hinders a high resolution clustering of genes into functionally related gene group based on their expression profiles (Bozdech, et al., 2003). Continued data mining of the published *P. falciparum* transcriptome, in addition to new transcriptome studies of defined developmental

stages or mutant parasites or drug treatments, will provide a better understanding of the biology of the malaria parasite.

Large scale proteomic studies were recently established as powerful approaches to analyze global protein expression profiles, differential protein expression, posttranscriptional posttranslational regulation and modifications, alternative splicing and processing, subcellular localization and interactions of host and pathogen (de Hoog and Mann, 2004; Zhang, et al., 2005; Zhu, et al., 2003). Several detailed high-throughput mass-spectrometry studies of the *P. falciparum* proteome have been published (Florens, et al., 2002; Khan, et al., 2005; Lasonder, et al., 2002; Le Roch, et al., 2003). It was shown, in general, the protein profiles agree well with their transcriptional profiles of the genes encoding these proteins but in many cases there is a slight delay between transcript production and protein accumulation (Le Roch, et al., 2003). Nevertheless many gaps remain in our understanding of protein expression. Detailed analyses of the *P. falciparum* proteome, and its relationship to the transcriptome would considerably benefit the annotation of the genome and functional genomics applied to the lifecycles of *Plasmodium*.

Understanding the interactions between *Plasmodium* proteins can provide insights into the function of many proteins as well as functional relationships with molecular mechanisms in the *Plasmodium* cell. Recently, the first large-scale analysis of interactions between proteins during the asexual blood stages of *P. falciparum* have been published (LaCount, et al., 2005). Using a high-throughput

yeast two-hybrid assay, 2,846 interactions were identified involving 1312 largely uncharacterized proteins in 29 highly connected protein complexes. Combining protein interactions with their gene expression profiles, putative annotation and domain information provided improved functional insights to *Plasmodium* biology, such as chromatin modification, transcription, mRNA stability, ubiquitination and invasion of host cells. Date and Stoeckert (2006) integrated the expression data and genomic context data (phylogenetic profiles and rosette stone data), available at that time, using naïve Bayesian method to construct an interactome of pair-wise functional linkages to elucidate local and global functional relationships between gene products (Date and Stoeckert, 2006). This resulted in predicting functional relationships between of 3667 proteins including 2216 hypothetical proteins at the 50% confident level. Wuchty and Ipsaro (2007) incorporated the evolutionarily conserved protein interactions, underlying domain-domain interaction information and experimental protein-protein interactions to construct a draft of protein interactions in malaria including only 2321 proteins (Wuchty and Ipsaro, 2007). Although two groups have developed methods to construct the malaria interactome separately to investigate the gene function, the resolution (confidence) and proteome coverage are not satisfying, especially for the network modular analysis and protein functional prediction. Hence, to construct one interactome with high confidence and proteome-coverage is urgent for the post-genomic research of malaria.

Taken together, the abovementioned large functional datasets are now available for *in silico* data mining. Integration of these heterogeneous functional data types and systematical analysis of these data are reliable and versatile to characterize the biology of the malaria parasite.

1.4 Systems biology for malaria

To study the transcriptome, proteome, interactome and other functional data integratively, bioinformatics tools are being developed to annotate the function of hypothetical proteins and point out specialized gene expression regulation systems in living organisms. The post-genomics research of *Plasmodium* species focuses on understanding of the transcriptome, proteome and interactome of the parasite to elucidate the gene regulation, cellular process, cell development. One of the main benefits from such research include understanding of the mode-of-action of inhibitory compounds which could explain resistance mechanisms to known drugs as well as identify and functionally describe new drug targets (figure 1.3) (Birkholtz, et al., 2006; Kooij, et al., 2006; Winzeler, 2006). Studies in model organisms suggest that most gene products mediate their function within complex networks of interconnected macromolecules. These networks have topological and dynamic properties that reflect biological phenomena. A comprehensive understanding of biological mechanisms associated with disease processes such as human malaria will require an interactome network whose confidence and gene coverage reaches the level on networks assembled for well studied model

organisms such as yeast, *C. elegans* or *Drosophila* (Barabasi and Oltvai, 2004). Only such networks can provide clues for the putative roles of pathogens genes in basic biological functions as well as adoption of pathogen cell to changing environments (Guo, et al., 2007). The functional predictions of unknown genes generated by such approaches are based on the gene connectivity, position in the network and other genes they have links with (Sharan, et al., 2007). Network-based predictions of protein function using network modular analysis and computational methods is presently one of the most powerful methods to predict the functions of the uncharacterized genes (Chua, et al., 2006; Deng, et al., 2003; Hishigaki, et al., 2001; Schlitt, et al., 2003; Schwikowski, et al., 2000; Sharan, et al., 2007). Although several data types of genome, proteome, transcriptome and interactome are available, malaria research is still in a period of intense data collection. Thus it is necessary and crucial to produce large functional data ensuring to provide

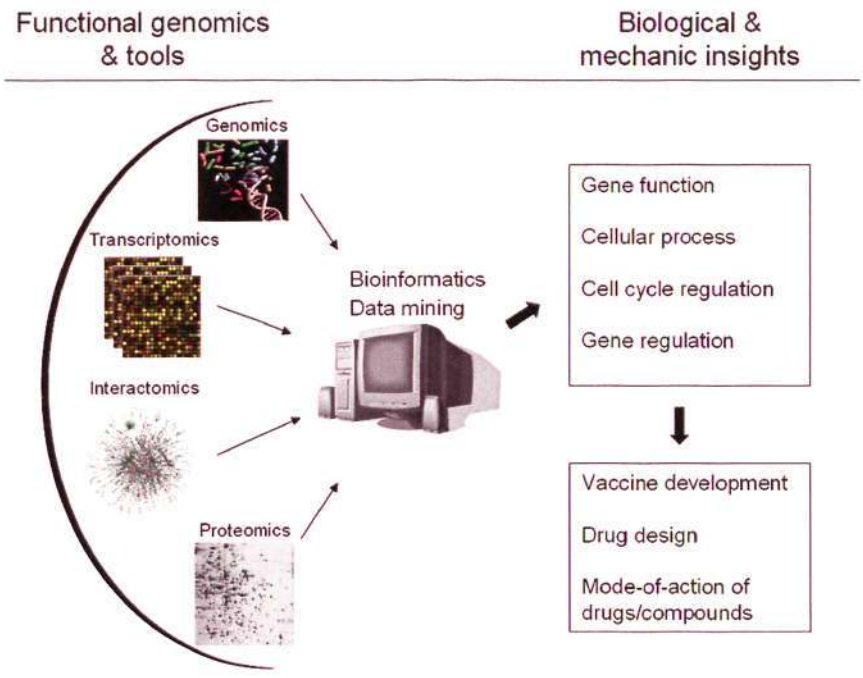


Figure 1.3 Functional genomics and systematic researches of malaria biology.

enough information for each gene and protein of *P. falciparum*. Integration and systematically analysis of these data types could facilitate to understand the *Plasmodium* biology with gene function and molecular processes and develop new drugs and vaccines.

1.5 Malaria merozoite invasion process

Merozoite invasion is a complex, multiple-step process in which four distinct steps can be recognized: 1) Initial recognition and attachment, 2) Reorientation, 3) Junction formation, 4) Parasite entry (figure 1.4) (Chitnis and Blackman, 2000; Dowse and Soldati, 2004; Pinder, et al., 2000). The major mission of the blood stage merozoites is to locate, bind to and invade host RBCs. Invasion is initiated by interaction between any part of the merozoite surface and the host cell. This initial interaction is likely a random collision and appears to be low affinity and reversible (Blackman, 2000). The cell recognition and attachment processes are highly dependent on specific molecular interactions between parasite ligands on the merozoite and the host receptors on the erythrocyte membrane (Barnwell and Galinski, 1998). After binding to the erythrocyte, the parasite reorients itself toward host plasma membrane. The merozoite reorientation also coincides with a transient erythrocyte deformation (Aikawa, et al., 1978). After that a junction is formed between the apical end of the merozoite and erythrocyte membrane. During the invasion, three 'secretory' morphologically distinct organelles: micronemes, rhoptries and dense granules which are located at the apical end of the invasive

stages of the parasite, expel their contents from the parasite immediately after the junction is formed. Junction formation and microneme release occur at about the same time after which the rhoptries are discharged immediately (Sam-Yellowe, et al., 1988). The release of the apical organelle contents correlates with the formation of the parasitophorous vacuole (PV). The apical location and the observation that the contents of the rhoptries and micronemes are released coinciding with invasion imply that these organelles participate in the invasion process. Moreover, the precise timing of the reorientation, organellar discharge, and formation of the PV suggest a tight regulation of these processes which are essential for a successful completion of the invasion process. Presently, close to nothing is known about molecular mechanisms associated with this tight regulation. As the merozoite moves through the ring-shaped tight junction formed by the receptor/ligand complex to the posterior of merozoite, the junction between the parasite and host becomes ring-like and the incipient parasitophorous vacuole (PV) is being formed. Once the parasite has completed its entry, the tight junction will disappear, and the respective parasitophorous vacuole membrane (PVM) and the host erythrocyte membrane will separate. The closure of the PVM is followed by the release of dense granule content into the lumen of PV. It is believed that merozoite invasion process involves complex machinery comprised of a broad spectrum of *Plasmodium* proteins. Several protein categories were linked with invasion, such as adhesive surface molecules, proteins involved in recognition and attachment, proteases essential for parasite and host protein modification and

degeneration, components of the actin-myosin motor complex, two type protein kinases (Blackman and Bannister, 2001; Cowman and Crabb, 2002; Dowse and Soldati, 2004; Preiser, et al., 2000). Despite these achievements, specific roles of the majority of the identified proteins in the invasion process are largely unknown. Identification of additional proteins associated with invasion as well as comprehensive understanding of their role in the invasion process is of outmost interest. First, comprehensive characterizations of the spectrum of merozoite surface molecules will be invaluable for vaccine development. Second, exploration of the unique molecular processes associated with the formation, regulation of the merozoite invasion machinery can provide many insights for drug development.

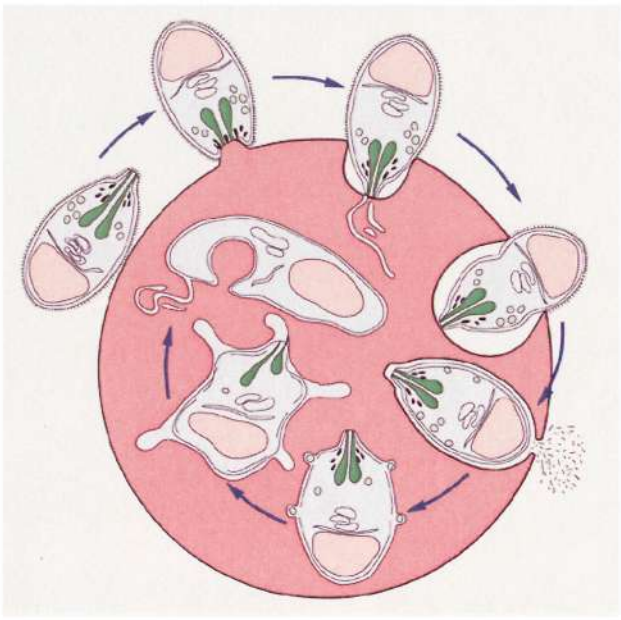


Figure 1.4 A schematic depiction of stages in red blood cell invasion by the malaria merozoite. The parasite binds, reorientates until its apical end contacts the host cell surface, then enters into a parasitophorous vacuole (figure was copied from (Chitnis and Blackman, 2000).

1.6 Project summary

Techniques of system biology have shown to significantly contribute to infer biological functions for the high number of uncharacterized proteins and to the understanding of biological mechanisms associated with disease processes. The fact that more than 50% of the genes of *P. falciparum* are of unknown function promoted us to characterize the function of these genes using system biology approaches. The publicly available transcriptional data characterizing of transcriptional regulation during the *P. falciparum* IDC that have been used in previous bioinformatics approaches for functional analyses brought only a limited contribution. This is mainly due to the monotonous character of the transcriptional regulation where many functionally unrelated gene share common transcriptional profiles. Given these limitations, growth perturbation data were suggested to be helpful for *Plasmodium* systems biology approaches. However, until today, very little is known about transcriptional responses of *P. falciparum* to growth perturbations. Here I propose to carry out growth perturbations of *P. falciparum* exposure to anti-malarial drugs and compounds. The generated gene expression data will be used to construct a gene-associated network by combining this dataset with other high throughput genomic datasets, such as phylogenetic profiles, domain-domain interactions and yeast two-hybrid protein-protein interactions. This network will be used to assign function to unknown proteins. In the final step I will focus on the Plasmodium invasion machinery and construct a gene sub-network to identify new proteins associated with cellular process and illustrate the molecular

mechanisms of this process. New identified proteins will be characterized using several methods of molecular biology. My work is presented with the following chapters: Chapter 2, develop a program for oligonucleotide selection for microarray and design one high quality DNA microarray of *P. falciparum*. And also several other DNA micorarrays are designed. Chapter 3, perform perturbations of *P. falciparum* exposure to 20 antimalarial drugs and compounds; analysis of the perturbation data; reconstruct a probabilistic gene functional network with the perturbation date combining with other high throughput functional datasets; network-based gene annotations; build a subnetwork associated with invasion process; experimental validation the newly identified invasion proteins. Chapter 4, transcriptional profiling *P. falciparum* exposure to kinase inhibitors. Chapter 5, the computing structure and databases developed in this project. Chapter 6, final summary and perspective.

Chapter 2 OligoRankPick: long oligonucleotides selection for DNA microarrays using weighted rank-sum strategy

2.1 Summary

The design of long oligonucleotides for spotted DNA microarrays requires detailed attention to ensure their optimal performance in the hybridization process. The main challenge is to select an optimal oligonucleotide element that represents each genetic locus/gene in the genome and is unique, devoid of internal structures and repetitive sequences and its Tm is uniform with all other elements on the microarray. Currently, all of the publicly available programs for DNA long oligonucleotide microarray selection utilize various combinations of cutoffs in which each parameter (uniqueness, Tm, and secondary structure) is evaluated and filtered individually. The use of the cutoffs can, however, lead to information loss and to selection of suboptimal oligonucleotides, especially for genomes with extreme distribution of the GC content, a large proportion of repetitive sequences or the presence of large gene families with highly homologous members. In this work we present the program OligoRankPick which is using a weighted rank-based strategy to select microarray oligonucleotide elements via an integer weighted linear function. OligoRankPick is an efficient tool for the design of long oligonucleotide DNA microarrays which does not rely on direct oligonucleotide exclusion by parameter cutoffs but instead optimizes all

parameters in context of each other. OligoRankPick provides significant improvements in oligonucleotide design in comparison to other published algorithms for all three testing microbial genomes *E. coli*, *S. cerevisiae* and *P. falciparum*. The weighted rank-sum strategy utilized by this algorithm provides high flexibility of oligonucleotide selection which accommodates extreme variability of DNA sequence properties along genomes of many organisms. Also we showed applications of OligoRankPick to design DNA microarrays for other species.

2.2 Introduction

DNA microarray is one of the most powerful and versatile tools for post-genomic research (Brown and Botstein, 1999). After the initial success with cDNA and PCR product-based microarrays, application of long oligonucleotides became widely used in "spotted" DNA microarray technology in the last eight years (Bozdech, et al., 2003; Hughes, et al., 2001; Kane, et al., 2000; Li and Stormo, 2001). From the beginning it became clear that the design of the oligonucleotide probes requires special attention. Under a single stringency condition, hybridization specificity and efficiency of all oligonucleotides must be globally maximized across the entire array. Thus for the selection of the optimal oligonucleotide candidates, four major parameters are being evaluated: (i) uniqueness which analyzes other possible cross-hybridization targets in the genome; (ii) sequence complexity which evaluates the presence of short nucleotide repeats; (iii) melting temperature (Tm)

or GC content which ensures a uniform hybridization efficiency across the microarray; and (iv) level of internal secondary structures which helps to avoid all possible self-binding interference with the specific target hybridization. In principle each of these properties can be calculated individually for every potential oligonucleotide candidate, however, the main challenge that remains is to derive a selection strategy that combines these parameters and selects the most optimal oligonucleotide representative for a given genetic locus/gene.

All currently available programs for long oligonucleotide microarray design utilize different parameters: the binding energy or BLAST-based score to alternative targets to evaluate uniqueness, the GC content or Tm to estimate hybridization stringency, the reverse Smith-Waterman score or free energy to evaluate levels of secondary structure and various types of complexity coefficients to evaluate the presence of short nucleotide repeats in each oligonucleotide element (Bozdech, et al., 2003; Nielsen, et al., 2003; Nordberg, 2005; Reymond, et al., 2004; Rouillard, et al., 2003; Wang and Seed, 2003; Wright and Church, 2002). Typically these programs select one or more oligonucleotide representatives of a gene using various systems of cutoff-based filters. For example ArrayOligoSelector creates an intersection of oligonucleotides that pass parameter-based cutoffs for uniqueness, self-binding and sequence complexity. The intersection candidate list is then passed on to the GC filter and subsequently the final representative(s) are selected using a 3' proximity criteria (Bozdech, et al., 2003). The cutoff based algorithms provide a powerful approach to select DNA

microarray oligonucleotide sets and were successfully used to design DNA microarrays for a large number of species (Boyer, et al., 2005; Bozdech, et al., 2003; Carter, et al., 2005; Nordberg, 2005). The use of these algorithms is, however, not completely optimal for genomes with high abundance of repetitive sequences and large fluctuations of GC content. To accommodate such genomic sequences, the methods must relax the parameter filter adjustments. The wide "opening" of the cutoff filters can cause selection of suboptimal oligonucleotides for a significant number of genes, due to the fact that all oligonucleotides that pass a particular filter are treated as equal by the subsequent steps, disregarding their subtle diversity within the filtered interval of the parameter (unpublished observations).

To overcome these shortcomings new algorithms which incorporate optimization strategies of oligonucleotide parameters were developed including OligoDesign (Tolstrup, et al., 2003) and CommOligo (Li, et al., 2005). OligoDesign was developed specifically for the design of the locked nucleic acid (LNA) microarray platform which takes advantage of the improved nucleic on-chip capture sensitivity of the LNA substitute mixmer oligonucleotides. Design of these specialized probes requires careful optimizations of the hybridization specificity and efficiency for each probe. For this purpose, OligoDesign uses an extensive fuzzification process derived from neural network approaches to ensure the optimal performance of this highly specialized microarray platform (Tolstrup, et al., 2003). Similar to the fuzzy logic approach,

CommOligo uses a piece-wise linear function to select optimal oligonucleotides via a user configurable iterative process (Li, et al., 2005). Both of these methods represented a step in the right direction, recognizing the need for parallel optimization of all used parameters and elimination of cutoffs that cause information loss. At its presently available implementation, however, both OligoDesign and CommOligo utilize complex and computer-time consuming processes that render them unsuitable for high throughput applications. Nevertheless both methods have been useful for design of focused "miniarrays" which typically contain smaller numbers of genes *e.g.* 120 stress response and toxicological markers from *Caenorhabditis elegans* (Tolstrup, et al., 2003) or microarrays for relatively small genomes such as *Methanoccocus maripaludis* with 1759 genes (Li, et al., 2005).

We developed a novel program named OligoRankPick (Hu, et al., 2007) that it is suitable for the design of gene specific long oligonucleotide probes for genomes of all sizes and the final decision making process is based on a weighted rank-sum strategy for parameter optimization which significantly streamlines the entire computation process. This program completely eliminates all cutoff-based filters, thereby significantly improves the quality of the resulting microarray oligonucleotide design. Moreover, the weighted rank-sum approach enables users to implement an integer weighted linear function to automatically optimize the oligonucleotide parameters for each gene individually.

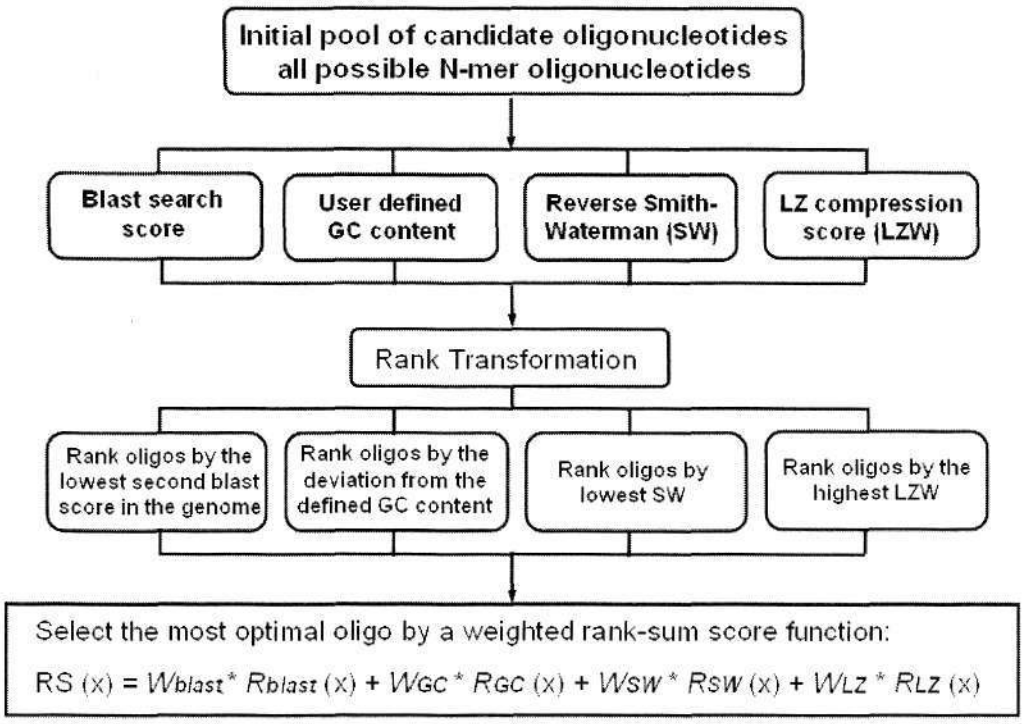


Figure 2.1. The flowchart of OligoRankPick. All possible oligonucleotides were extracted form the input sequence and stored. Subsequently four parameters of all possible oligonucleotides were calculated including the BLAST score to a second genomic target (uniqueness), the GC content (Tm), the Reverse Smith-Waterman score (self-binding) and the LZ compression score (sequence complexity). In the rank transformation step, the oligonucleotides are ranked based on each parameter and ordinal rank number is given to all oligonucleotides in each parameter rank independently. Finally weighted rank-sum ($RS_{(x)}$) is calculated for all oligonucleotides with uniqueness weights (W_{BLAST}), GC content weights (W_{GC}) self-binding weights (W_{SW}), and sequence complexity weights (W_{LZ}) and R_{BLAST}, R_{GC}, R_{SR} and R_{LZ} representing the ranks corresponding to each parameter ranking. Multiple $RS_{(x)}$ are determined by the gene specific optimization using multiple weight sets (not indicated) and the lowest value is finally considered. The

optimal candidate is selected based on the lowest $RS_{(x)}$ amongst all oligonucleotides in the locus.

2.3 OligoRankPick

2.3.1 Program overview

Figure 2.1 summarizes the global overview of the OligoRankPick algorithm. Essentially, all possible oligonucleotide windows from a gene/locus are extracted and scored by the four parameter measurements, uniqueness (BLAST score to second target), GC content (GC content, Tm), self-binding (Reverse Smith-Waterman, SW) and sequence complexity (Lempel-Ziv compression score) (figure 2.1). In the next step OligoRankPick ranks all possible oligonucleotides in one locus according to their parameter scores and assigns an ordinal number for each parameter. While the BLAST, SW, LZ score are directly transformed into a rank, the GC content scores are first transformed to their absolute deviation from the defined GC content. Oligonucleotides with an identical score for any parameter are offered the same rank number. Subsequently the rank-sum strategy is used to select the optimal oligonucleotide(s). This strategy is based on the calculation of a weighted rank-sum of all four ranks for each oligonucleotide within a locus by a linear function utilizing the following formula (also see figure 2.1):

$$RS_k = \min_k\left(\sum_{j=1}^{4} W_j * R_{jk}\right)$$

Where W_j is the weight of the j-th parameter (j = 1, 2, 3, 4), R_{jk} is rank score of j-th parameter of the k-th oligonucleotide (k = 1, ..., n). In the first step the rank-sum

function selects the oligonucleotide with the minimal rank-sum (RS) as the candidate for one given weight set.

To accommodate the variable characteristics of the DNA sequence along the genome we introduce an additional step in which the optimal weight values are determined for each gene individually. There is a weight file (wt_pool.opt) to offer the optimal intervals of weight values for the user from the simulation for "standard" microbial genomes, which is detailed in the published paper (Hu, et al., 2007). However, users can define specific weights and modify this file based on their own theoretical or empirical experience as well as specific requirements (simulation_ws.pl provided in the package). For all sets of weights:

$$TO = \min_{i}(RS_{Ki} / \sum w_i)$$

Where RS_{Ki} is the optimal selected oligonucleotide (K oligonucleotide) for weight set i, Σw_i is the sum of weights for weight set i. TO (Target Oligonucleotide) is the final selected oligonucleotide. The optimization step is performed for all weight sets reflecting all combination of weight values in the input intervals. The oligonucleotide with the minimum RS value is the optimal local solution of the rank-sum function in the given weight set interval. This oligonucleotide is chosen as the final candidate.

2.3.2 Implementation

The OligoRankPick program is freely available from the website (zblab.sbs.ntu.edu.sg/OligoRankPick). It is divided into two parts (two scripts).

The first script (oligoblast.pl) is used to generate all possible oligonucleotides and their parsed BLAST results including its first, second and third best hybridization target (top three). The oligoblast.pl script can be run on different computers or a computer cluster using parallel processing methods such as mpiBLAST (www.mpiblast.org) and the results should be parsed according to the format of oligoblast.pl output. The second script (oligorankpick.pl) selects the optimal oligonucleotide for each sequence. There are four additional scripts which can be used to optimize the OligoRankPick package performance including masker.pl, used to mask the repeat sequence based on the NCBI dust program; GC_dis.pl, used to plot the GC content distribution of all oligonucleotides in the dataset in order to define a suitable GC content; fragmentation.pl, used to partition the long sequences to increase the oligonucleotide density in the coding sequences (see *P. falciparum* microarray design); simulation_ws.pl, used to modify the weight set file (wt_pool.opt) for special genomes.

2.3.3 Comparison with other programs

To compare the performance of OligoRankPick with other publicly available programs, we designed three theoretical microarray oligonucleotide sets for the *P. falciparum*, *S. cerevisiae* and *E. coli*. We selected three programs, ArrayOligoSelector (Bozdech, et al., 2003), OligoPicker (Wang and Seed, 2003) and OligoArray 2.1 (Rouillard, et al., 2003). For the intended designs we chose the oligonucleotide length to be 70 nt and the GC content 31.4% (Tm=74.7) for *P.*

falciparum, 40% (Tm=79.8) for *S. cerevisiae* and 45% for *E. coli* (Tm=82.7). The theoretical oligonucleotide sets were designed using the publicly available sequence data and the selection algorithms with default settings. Figure 2.2 summarizes the parameter distributions of the uniqueness scores (BLAST scores of the final oligonucleotides to their second best genomic targets) plotted against GC content. Overall these contour plots illustrate that comparing to the three publicly available programs, OligoRankPick provides significant improvements for the design of yeast, *E. coli* and *P. falciparum* microarray (figure 2.2). The most striking improvements were, however, observed in the design of the *P. falciparum* microarray. For this genome the BLAST scores and the GC content of the oligonucleotides designed by OligoRankPick exhibit a greater convergence to a small region in the desired area (low BLAST scores, GC around 31.4%) compared to oligonucleotides designed by the three other programs (figure 2.2).

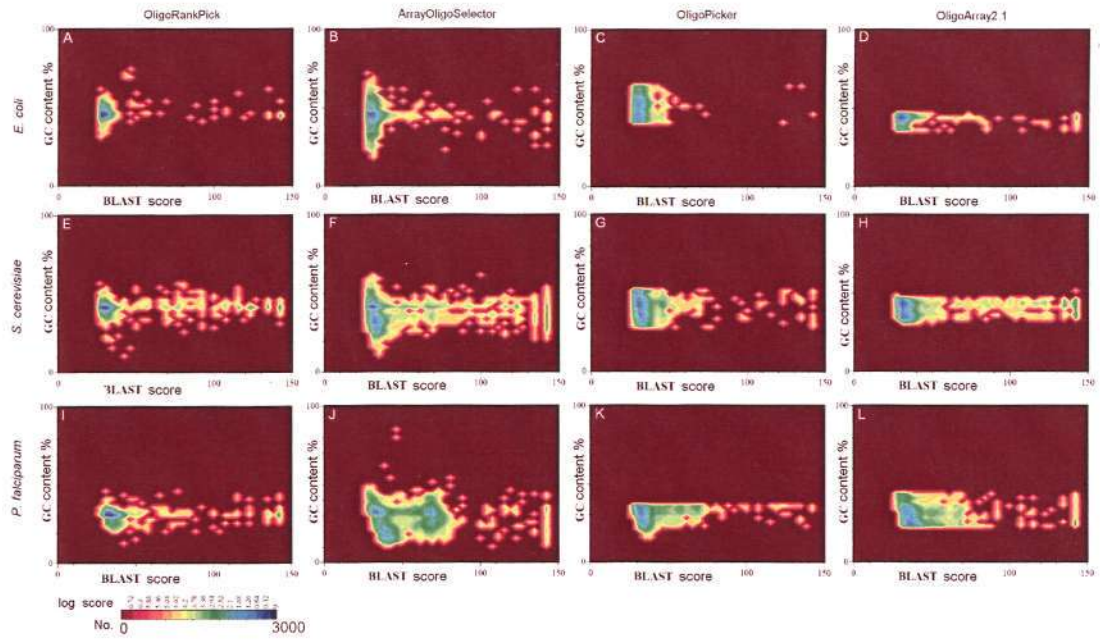


Figure 2.2 Overall profiles of the uniqueness and GC content of oligonucleotide

microarray elements in the 12 designed theoretical microarray sets. Four algorithms OligoRankPick, ArrayoligoSelector, OligoPicker, and OligoArray2.1 were used to design long oligonucleotide DNA microarray sets for *P. falciparum*, *E. coli* and *S. cerevisiae*. Contour plots illustrate oligonucleotide density plotted of along the uniqueness scores (second target BLAST scores) and GC contents. The oligonucleotide density is calculated as $-\log10(N/N_{max})$ (N ~ number of oligonucletide in a given area and N_{max} ~ number of oligonucleotide in the most dense area) and displayed using the indicated by the color based scale.

Similar convergence is observed for the SW and LZ scores (Supplementary figure S2.1 and S2.2). To further demonstrate the convergence of the oligonucleotide parameters we calculated a mean distance for each parameter distribution to its desired (preset) value and also to the average value within the parameter distribution (figure 2.3). In all cases the OligoRankPick produced the smallest mean distances and thus tighter distribution of the oligonucleotide parameters. The only exception is the lower mean distance of the CG content from its mean value in the yeast set designed by OligoPicker. Detailed inspection of these results indicated that the low mean distance is due to extensive filtering implemented by this program (data not shown). For each of the theoretical microarray dataset we also calculate the average weight score (AWS) which is directly related to the oligonucleotide quality with respect to the oligonucleotide parameters. The smaller AWS that are consistently observed for the

OligoRankPick generated oligonucleotide sets compared to the three other programs further indicate the optimization power of OligoRankPick (Supplementary figure S2.3).

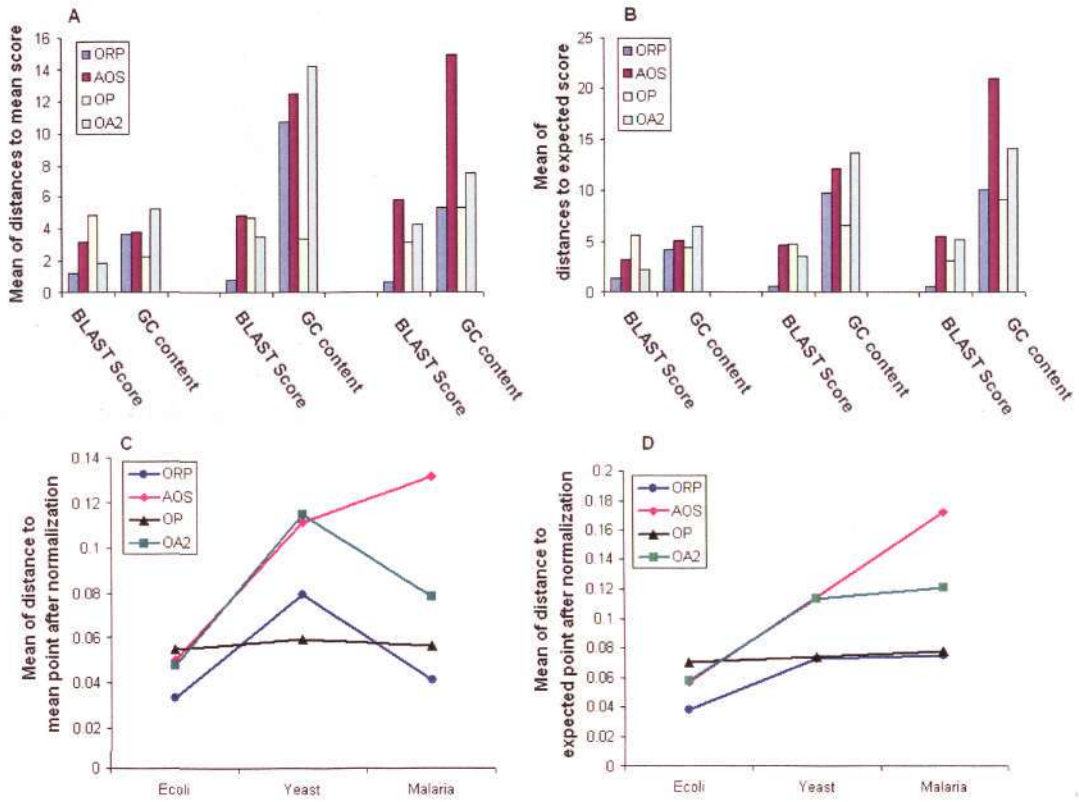


Figure 2.3 Analyses of the uniqueness and GC content distributions in the 12 designed theoretical microarray sets. A, single-parameter mean distances of BLAST score and GC content were calculated from all oligonucleotide scores to their mean score, respectively; B, the mean distances of BLAST score and GC content calculated from all oligonucleotide scores to the expected score. The expected BLAST score is the smallest one in all sets and the expected GC content is the defined GC content in the program; C and D shows the two-dimensional mean distances BLAST score and GC content calculated from all oligonucleotide

points to their central point. The central point in C comprises the mean BLAST score and mean GC content in the oligonucleotide set. The central point in D has the smallest BLAST score and defined GC content.

Table 2.1 summarizes the overall statistics of the 12 oligonucleotide sets for different datasets and methods. We define the 40% continuous sequence identity to a second target and 5% deviation from the target GC content as the "good quality" criteria according to previous studies (Bozdech, et al., 2003; He, et al., 2005; Hu, et al., 2007; Kane, et al., 2000). OligoRankPick outperformed the other programs producing the highest number of oligonucleotides within the target limits (95.6%, 91.3% and 94.9% for *E. coli*, *S. cerevisiae* and *P. falciparum* respectively, table 2.1). The unbiased character of the OligoRankPick algorithm is also demonstrated by the total number of oligonucleotides designed. Since OligoRankPick does not use any filters, this method will select an oligonucleotide candidate for essentially any genetic locus (see "#designed" in table 1). There were only 5 coding sequences not considered by OligoRankPick in *E. coli* and one in *S. cerevisiae* due to their sequence lengths being shorter than 70 nt (table 2.1).

One of the unique features of the *P. falciparum* genome is the presence of several large highly homologous gene families whose role has been implicated in the antigenic variation including *var* (76 members), *rifin* (164 members) and *stevor* (34 members) (Florens, et al., 2002; Kyes, et al., 2001). Table 2.2 indicates the number of unique oligonucleotides designed by all the four programs

for these genes. OligoRankPick was capable of designing unique oligonucleotides for 234 genes (85.4%) of total 274 genes which by far exceeded the performance of the three other algorithms. Analysis of oligonucleotide positions of *var* genes showed that they located at the most variable regions such as the internal sequence (ITS) between two conserved domains (figure 2.4).

Table 2.1 The comparison of designed oligonucleotides from different programs

Programs*	*E. coli* K12 (4237 cds)		*S. cerevisiae* (6680 cds)		*P. falciparum* (5363 cds)	
	#designed[a]	#accepted [b]	#designed	#accepted	#designed	#accepted
OligoRankPick	**4232**[ξ]	**4047(95.6)** [&]	**6679**[ζ]	**6096(91.3)**	**5363**	**5092(94.9)**
ArrayOligoSelector	4201	3371(80.2)	6221	3471(55.8)	5339	2093(39.2)
OligoPicker	4142	2594(62.6)	6208	3614(58.2)	4235	3543(83.7)
OligoAarray 2.1	3221	2826(87.7)	6587	4440(67.4)	5206	2317(44.5)

*ArrayOligoSelector 3.8.4; OligoPicker; OligoArray 2.1. a: oligonucleotide number selected by the program; b: good oligonucleotide number based on BLAST score of non-target (<=40% continuous identity) and GC content (±5%). & Percentage of good quality oligonucleotide to total selected oligonucleotide (in the bracket). ξ Five rejected coding sequences are less than 70bp. ζ Only one rejected sequence is YJR151W-A (51bp).

Table 2.2 The oligonulceotide design of large gene families from different programs

Programs*	*Var family* (Total No. 76)		*Rifin family* (Total No. 164)		*Stevor family* (Total No. 34)	
	#designed[a]	#accepted[b]	#designed	#accepted	#designed	#accepted
OligoRankPick	**76**	**63**	**164**	**140**	**34**	**31**
ArrayOligoSelector	76	31	162	58	34	13
OligoPicker	37	37	78	74	12	12
OligoAarray 2.1	22	9	162	118	34	22

*ArrayOligoSelector 3.8.4; OligoPicker; OligoArray 2.1. a: oligonucleotide number selected by the program; b: accepted oligonucleotide number based on BLAST score of non-target (<40% continuous identity).

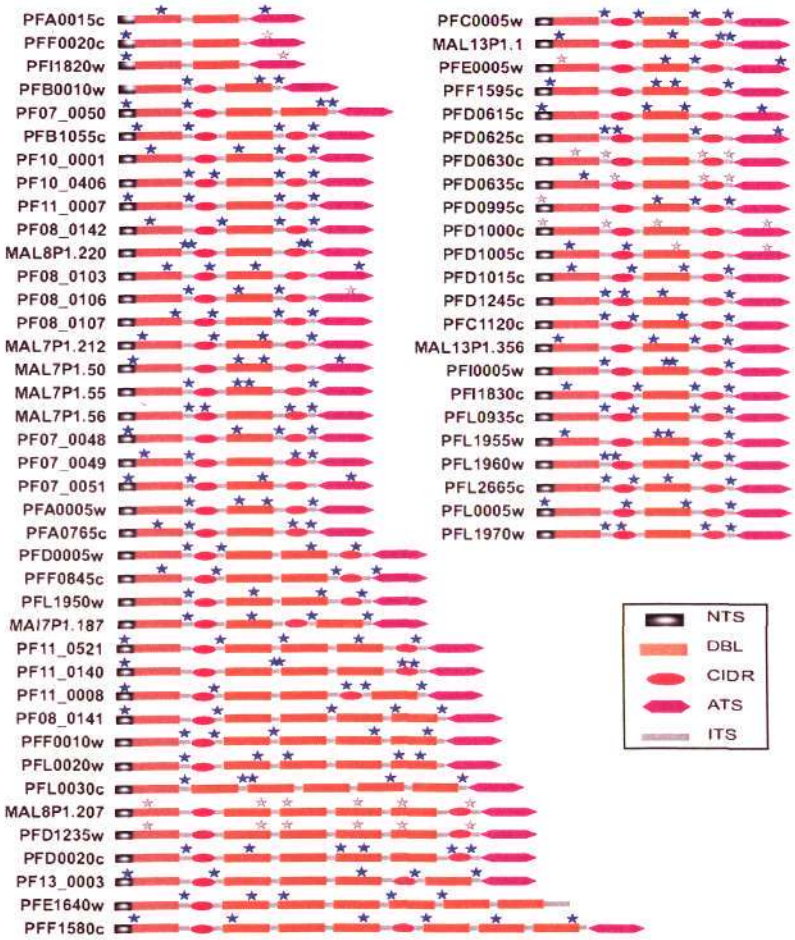


Figure 2.4 The positions of selected oligonucleotides by OligoRankPick for *var* genes (not including pseudogenes). Oligonucleotide position was marked by blue star (unique) and gray star (nonunique). NTS, N-terminal sequence; DBL, Duffy-binding domain; CIDR, cysteine-rich interdomain region; ATS, acidic terminal segment; ITS, internal sequence.

2.4 DNA microarray for *Plasmodium falciparum*

2.4.1 Design of a gene specific DNA microarray for *P. falciparum*

In the final step we applied OligoRankPick to design a gene specific DNA microarray for the *P. falciparum* genome (5363 coding sequences, CDS) that can be used for functional genomic studies of this important human pathogen. For this design we wished to increase the oligonucleotide coverage for longer open reading frames and thus we fragmented each coding sequence using the fragmentation.pl script as follows: sequences smaller than 1kb were kept as one fragment; sequences between 1kb and 2kb were split evenly into two fragments, sequences larger then 2kb were split into n fragments (n>=2) when: (2n-2)kb < gene size > (2n)kb. The fragmentation step generated 10166 Microarray Element Fragments (MEFs) from 5363 CDS. A single oligonucleotide was designed for each MEF which resulted in one oligonucleotide per 1198bp on average for all *P. falciparum* coding sequences. Although the median GC content of all 70 nt oligonucleotide windows in the *P. falciparum* coding sequences is 24.3% (displayed by GC_dis.pl optional module) for higher specificity and efficiency of microarray hybridization, we selected oligonucleotides with a GC content of 31.4% (22 GCs out of 70 nt). OligoRankPick successfully designed 10166 oligonucleotides representing all predicted *P. falciparum* genes with an average of 1.9 oligonucleotides per protein coding sequence. Figure 2.5B summarizes the GC content distribution suggesting that OligoRankPick can identify optimal oligonucleotide elements with GC content significantly distant from the average GC content in the genome. Astonishingly 70.5% of the designed oligonucleotides had the desired GC content

of 31.4% (figure 2.5B).

To evaluate the level of uniqueness of the designed oligonucleotides we used the identical quality control criteria used for the weight optimization strategy which is consistent with previously established conditions of optimal microarray hybridization performance (see above). In total 9909 (97.5%) oligonucleotides passed the uniqueness criteria and 9795(96.4%) oligonucleotides were found to be in the range of 5% deviation from the GC content target value (31.4%) (figure 2.5). There are 9584 (94.7%) oligonucleotides meeting both criteria while only 275 oligonucleotides (2.7%) were outside of the ±5% GC content interval and 257 oligonucleotides (2.5%) were not unique in the genome. Manual inspections of the MEFs represented by these oligonucleotides indicated that no suitable 70 nt window exists within these DNA fragments. The 257 non-unique oligonucleotides represented 193 genes (3.6% of total CDS) from which 67 genes belong to the large multigenic gene families, *var*, *rifin* and *stevor*. Pair-wise sequence homology analysis of these genes revealed that these genes do not contain any 70 nt window that shares less than 40% homology with any other member of the corresponding gene family and thus no unique oligonucleotide could be selected by any conceivable strategy. Interestingly for the remaining 185 (73.4%) members of these families a specific oligonucleotide was selected which further demonstrates the power of OligoRankPick for microarray design.

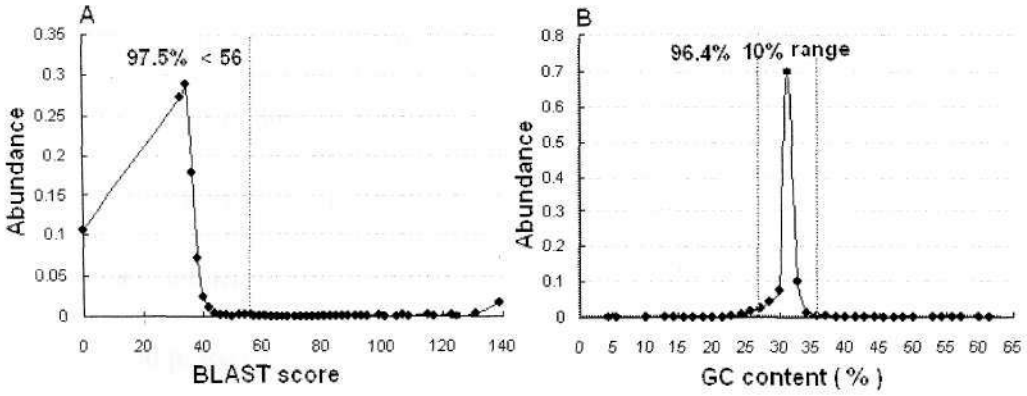


Figure 2.5 Oligonucleotide parameter distributions in the newly designed *P. falciparum* DNA microarray. Total 10166 oligonucleotides were designed for the P. falciprum DNA microarray. Relative abundance of the oligonucleotides is plotted along the uniqueness scores (BLAST score of the second-best target in the genome) (A) and along the GC content (B). The dotted line indicates the quality control criteria (see text) with BLAST score = 56 which corresponding to >40% continuous match cross-hybridization and the 31.4% ± 5% interval of GC content corresponding to the targeted range. Percentages of oligonucleotides which fall within the targeted values are indicated.

2.4.2 Transcriptome analysis of the trophozoite and schizont stages of *P. falciparum*

Although all parameters of the oligonucleotide microarray sets designed by OligoRankPick indicate their high quality, the ultimate evidence for their functionality can be provided only by physical microarray experiments. For this purpose we have synthesized all the 10166 oligonucleotides for the *P. falciparum*

genome-wide microarray and spotted these onto polylysine-coated microscopic slides as previously described (DeRisi, et al., 1997). Using these microarrays we compare the global mRNA patterns between two developmental stages of the *P. falciparum* intraerythrocytic development, trophozoite and schizont. All experimental procedures were carried out as previously described (Bozdech, et al., 2003) and the complete results for three replicates of the microarray hybridizations are available in the supplementary data. The *P. falciparum* genome sequence reference strain 3D7 was used for this analysis. Total 4183 genes were found to be expressed in at least one of the studied developmental stages in three replicates of microarray hybridization. From these 1891 and 841 mRNA transcripts exhibited at least 2-fold higher abundance in the trophozoite and the schizont stage, respectively.

Table 2.3 *P. falciparum* microarray data and their comparisons to existing transcriptomes

Transcriptome results	**Trophozoite**	**Schizont**
3-fold in at least two replicates	**862**	**431**
Present in the LOM-IDC transcriptome	630/73%	320/74.2%
*Same stage classification in LOM-IDC Transcriptome	595/94.5%	307/95.9%
Present in the HDSO-Affymetrix transcriptome	741/86%	353/82%
**Same stage classification in HDSO-Affymetrix transcriptome	676/91.2%	336/95.2%

*genes with peak expression before and after 30 hours post invasion are classified as trophozoite and schizont specific, respectively.

** genes with higher expression levels in late ring and early and late trophozoites

compared to early and late schizonts are classified as trophozoite specific and *vice versa.*

In order to assess the fidelity of the obtained results we wish to compare this data to previously published transcriptome analyses of the *P. falciparum* intraerythrocytic developmental cycle (IDC). These include the IDC transcriptome analyzed by the previous version of a long oligonulceotide microarray (LOM-IDC transcriptome) comprised of 2689 genes (Bozdech, et al., 2003), and a high density short oligonucleotide Affymetrix microarray dataset (HDSO-Affymetrix transcriptome) comprised of 1162 genes with stage specific transcription (Le Roch, et al., 2003). All genes present in both LOM-IDC and HDSO-Affymetrix transcriptomes were represented on the new *P. falciparum* microarray and yielded a hybridization signal in at least two of the three microarray replicates. To compare the stage specificity of the gene expression we select genes which exhibited >3-fold change in mRNA abundance between trophozoite and schizonts detected in at least two (out of three) replicates (table 2.4). Using these criteria we classify 862 genes as trophozoite specific and 431 genes as schizont specific. The transcriptome data comparisons, summarized in table 2.3, indicate high correlations between the transcriptome data and the new microarray dataset with 91.2-95.9% of overlapping genes exhibiting identical stage specificity in their mRNA levels. There were only a small number of genes (4.1-8.8%) for which the new expression results did not correlate with the previously published data.

These discrepancies are likely caused by subtle differences in parasite culture synchronicity and stage representation between our culturing system and the systems used for the previous transcriptome analyses.

To further validate the performance of the designed *P. falciparum* microarray quantitative real-time RT-PCR was used to measure relative mRNA abundance between trophozoite and schizont stage for 10 selected genes. For this we chose genes for which only OligoRankPick designed a "good quality" microarray element while the three tested publicly available programs did not yield a suitable oligonucleotide element. These include two paralogous histone3, five members of the variable surface antigen gene families (2 *var*, 1 *rifin*, 2 *stevor*), centrin, and two genes encoding highly homologous hypothetical proteins. Figure 2.6A shows good correlations between the RT-PCR results and microarray hybridization data which demonstrate the robust performance of the newly designed microarray for analyses of mRNA abundance in *P. falciparum*. Detail sequence analyses revealed that each of the 10 selected genes contains only a small window of unique sequence while the majority of the gene is highly homologous to at least one other locus in the genome. One of the example is a pair of highly homologous genes encoding histone3 (H3) and its homologue histone3.3 (H3.3) (figure 2.6B). This high homology is likely the main obstacle for designing a specific oligonucleotide and it is the reason why no transcription data have been obtained by the previously reported transcriptome analyses. Despite this OligoRankPick selected specific oligonucleotides which overlap the most unique region of each gene (figure 2.6B).

The microarray hybridization signal detected on these oligonucleotide elements revealed that these two highly homologous genes undergo different transcription regulation during the IDC with H3 exhibiting 3-fold increase of mRNA abundance in schizonts compare to trophozoites and H3.3 showing similar amounts (<2-fold change) between these two developmental stages (figure 2.6A).

Taken together these data demonstrate that the newly designed microarray for *P. falciparum* successfully recapitulates data from previous transcriptome analyses and has a potential to further expand on these results. Overall these data verify the improved performance of OligoRankPick in designing unique microarray elements for gene expression microarrays.

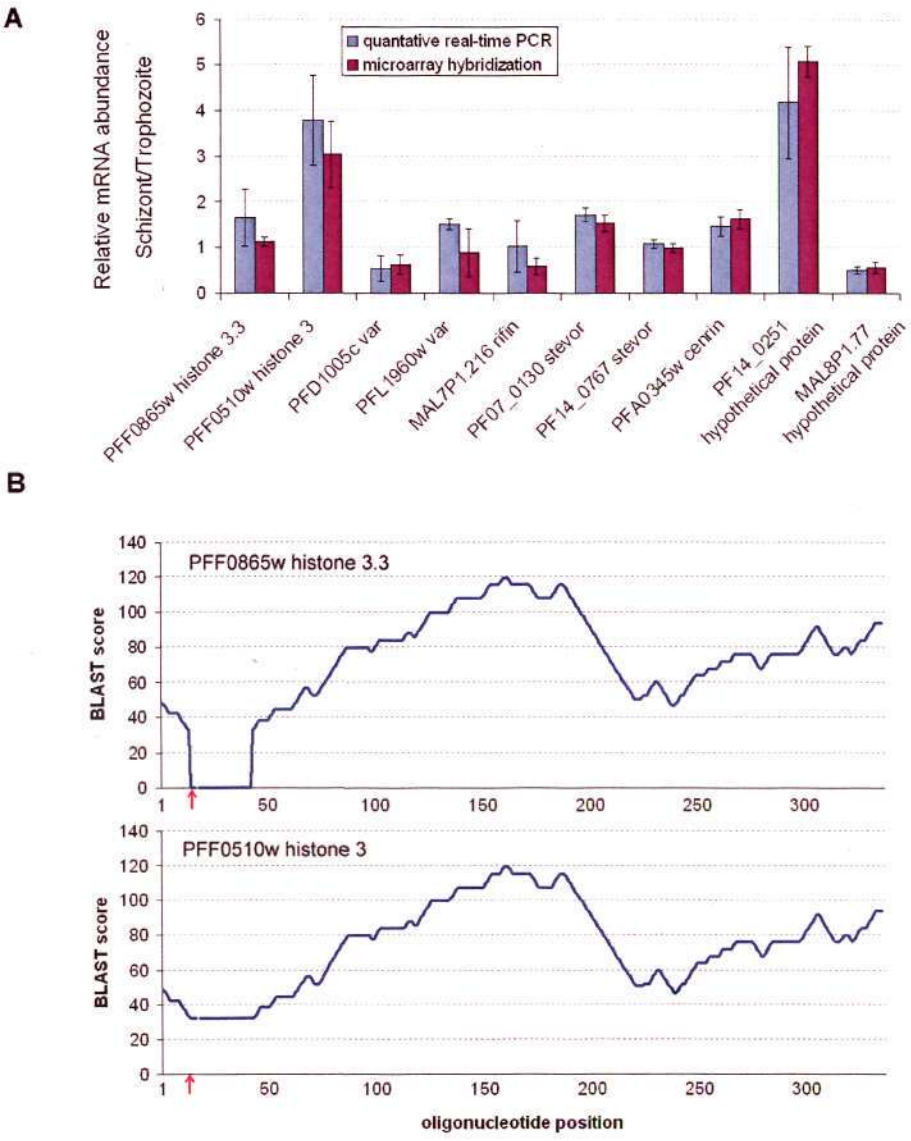


Figure 2.6 Verifications of microarray results by quantitative real-time PCR **(A)** and example of oligonucleotide selection for highly homologous genes **(B)**. The bar graph indicates mRNA abundance ratios between two developmental stages (schizont/trophozoite) of the *P. falciparum* IDC for 10 genes measured by microarray and by real-time RT-PCR. The expression data were obtained using the total RNA isolations from the trophozoite and schizont stage. Each measurement was carried three times and the standard error for each measurement is indicated. **(A)**. The uniqueness score distributions along the two highly homologous histone

3 genes. The uniqueness is represented by the BLAST score of each 70 nt window along the histone genes (H3 and H3.3) to its second best target in the genome. The red arrow indicates the position of oligonucleotide selected by OligoRankPick in each gene **(B)**.

2.5 Discussion

The main goal of this work was to develop a microarray design algorithm which combines the thoroughness of the parameter optimization methods (such as CommOligo (Li, et al., 2005)) and performs with high computational efficiency of the earlier, cutoff based techniques, such as OligoArraySelector (Bozdech, et al., 2003). The newly developed algorithm, OligoRankPick, is the first method using a parameter optimization approach that is computationally fast and robust for genome-wide microarray design. The core principle of this technique consists of the rank transformations of the parameter scores and the subsequent weighted rank-sum strategy. This allowed us to eliminate all cutoff based filters that are typically applied to the input data (by existing optimization programs) or to partial oligonucleotide lists that are generated prior or during the decision-making step (in cutoff-based methods). Instead the derived rank-based system maintains all the oligonucleotide candidates in their rank order throughout the entire process. This approach removes any ambiguities in the selection process as all oligonucleotides are constantly prioritized based on their properties. Since no oligonucleotides are eliminated by arbitrary cutoffs, this method also significantly expands the genome

coverage of the designed microarrays. The simplicity of the rank-based approach also allows the algorithm to perform gene specific optimizations of the weight coefficients in which the contribution of each parameter is modified based on the sequence properties of a particular gene. This is especially useful for optimal probe design in genes with extreme parameters distributions such as high AT content or high sequence homology to other genomic locus (low uniqueness). For example AT richness of some genes causes the GC content parameter to be over emphasized due to a stronger priority that is given to the GC rich oligonucleotide windows. This could force a selection of less unique oligonucleotides or oligonucleotides with complex secondary structure from these CG rich oligonucleotide candidates. The implementation of the gene specific optimizations is likely the most innovative approach introduced by this method because it generates a tighter distribution for each oligonucleotide parameter compared to other publicly available programs (figure 2.2). For general functionality we derived and validate optimal weight set intervals which could be applied to a wide range of genomes. The flexibility of the OligoRankPick package, however, allows the users to tune these setting for other specialized applications.

For the development and validation of OligoRankPick we design a new DNA microarray for the most lethal species of the human malaria parasites *P. falciparum* whose genome was completed in 2002 (Florens, et al., 2002). We chose this genome for its extreme AT/GC distribution and high level of gene duplication to demonstrate the utility of the newly design program for its future applications. The

average GC content in the *P. falciparum* genome is estimated to be 19.4% (23.7% in coding and 13.5% in non-coding sequences). For this design, however, we wished to select oligonucleotides with higher GC content to ensure higher Tm and thus specificity and selectivity of each probe. In addition the requirement for high GC content will help to select oligonucleotides with high sequence complexity as AT rich sequences in *P. falciparum* contain numerous short nucleotide repeats. As demonstrated in figure 3 OligoRankPick was able to design a set of oligonucleotides whose GC content is tightly distributed around 31.4%. At the same time high levels of uniqueness and sequence complexity and a low level of secondary structures were preserved in the vast majority of the probes. This feature of OligoRankPick will be particularly useful for microarray design of many organisms with extreme fluctuations in GC content such as *Mycoplasma mycoides* (Westberg, et al., 2004) and other bacterial species (Parkhill, et al., 2003), other "AT rich" *Plasmodium spp.* (Carlton, et al., 2002) and *Dictyostelium discoideum* (Glockner, et al., 2002) or GC rich *Leishmania spp.* (Ivens, et al., 2005). The *P. falciparum* genome was found to contain a large number of duplicated genes sharing high levels of homology (Florens, et al., 2002). The extreme examples are the three gene families (*var, rifin, stevor*) which are involved in the parasite virulence and are presently explored as potential molecular targets for malaria intervention strategies (Rowe and Kyes, 2004). Despite the high levels of homology amongst the individual members of these gene families, OligoRankPick was capable of designing specific oligonucleotide representative for 74.3% of these

genes which by far exceeded the performance of the three tested publicly available programs. This improved performance will render OligoRankPick useful for studies of many organisms with highly homologous, biologically significant gene families ranging from microbial pathogens (Stringer and Keely, 2001) to high eukaryotes (Harrison and Gerstein, 2002).

2.6 Applications of OligoRankPick for other species

2.6.1 Design an intergenic specific DNA microarray of *P. falciparum*

DNA microarray of intergenic regions is very useful for analysis of regulation of gene expression such as Chromatin Immunoprecipitations (ChIP) (Iyer, et al., 2001; Ren, et al., 2000). *P. falciparum* genome has high AT content, especially in the intergenic gene regions where the AT content readily exceeds 90%. Hence designing of DNA microarray for the gene intergenic regions could be challenging for essentially all oligonucletide selection programs. In our previous work OligoRankPick was shown to successfully select unique representative oligonucleotides even for genomic regions with extreme parameters such as AT rich sequences or low complexity regions. This is achieved via the automatic parameter optimization that does not rely on cutoff-based definitions (Hu, et al., 2007). Total 5411 UTR sequences were generated by extracting 1500 bp upstream sequences from the starting codons of the *P. falciparum* genes representing the intergenic region that contain both the untranslated RNA sequences and the promoter regions. In the absence of the mapping of

transcriptional start sites in the *P. falciparum* genome these regions are the closest estimation of the position of gene expression regulatory regions. In order to increase the intergenic microarray resolution, each UTR sequence was further fragmented into three 500bp long sequences, and one oligonucleotide for each fragment was designed by OligoRankPick. For the oligonucleotide nomenclature, each fragment were marked position -1, -2 and -3 according to its distance to the start codon. For ChIP-chip applications, 500 bp sequences from start codon ATG of 5411 genes were also generated to design oligonucleotides by OligoRankPick, and this oligonucleotide was marked position +1. For the design, we set oligonucleotide length at 50 nt and GC content at 28%. Prior to the application of the OligoRankPick we performed RepeatMasker.pl to masker all high repetitive AT regions in the UTR sequences. Total 14975 oligonucleotides were selected by OligoRankPick for 16233 fragmented UTR sequences with 5411 genes (table 2.4). Significantly, 95.1 % (5147) genes have at least one unique oligonucleotide, and 2890 genes have three unique oligonucleotides for all three fragments of UTR sequences. After the quality control filtering of the designed oligonucleotides based on both uniqueness and GC content deviation, 4944 genes (91.4) have at least one oligonucleotide representing their UTR sequences on the newly designed *P. falciparum* intergenic oligonucleotide microarray (table 2.4). In all 14975 oligonucleotides, 10791 (~72%) matched two preset criteria (40% similarity and 6% deviation of GC content), and 88.3 % oligonucleotides have their GC content matching the previously defined setting. This confirms that OligoRankPick could

accommodate extreme variability of DNA sequence properties to select a proper oligonucleotide that matches strict criteria for microarray performance (see above). These results show high percentage of oligonucleotides selected by OligoRankPick These results also show high percentage of the oligonucleotides selected by OligoRankPick for the intergenic regions of *P. falciparum* genes that have high uniqueness and low deviation of GC content. ChIP-chip analysis based on this micorarray is performed and preliminary results showed that this chip had good signal qualities (Chaal et al. manuscript in preparation).

Table 2.4 Statistics of intergenic specific DNA microarray of *P. falciparum*

	Designed (%) [$]	Unique (%) [@]	Unique & GC content (%) [*]
# genes have at least one oligo	5411 (100)	5147 (95.1)	4944(91.4)
# genes have -1, -2 and -3 oligos	3570 (66.0)	2890 (53.4)	1940 (35.9)
# genes have -1 and -2 oligos	4545 (84.0)	3292 (60.8)	2388 (44.1)
# genes have -1 and -3 oligos	4673 (86.4)	3494 (64.6)	2622 (48.5)
# genes have - 2 and -3 oligos	4668 (86.3)	3540 (65.4)	2777 (51.3)
# genes have -3 oligos	5097 (94.2)	4397 (81.3)	3924 (72.5)
# genes have -2 oligos	4923 (91.0)	4083 (75.5)	3446 (63.7)
# genes have -1 oligos	4955 (91.6)	4103 (75.8)	3421 (63.4)
# genes have +1 oligos	5403 (99.9)	5109 (94.4)	5039 (93.1)
Total oligos for UTR	**14975 (92.2)**	**12583 (84.0)**	**10791 (72.1)**
Total oligos for 4 positions	20378 (94.2)	17692 (86.8)	15830 (76.7)

[$] number oligonucleotides outputted by OligoRankPick is lower then expected (number in the bracket) due to the use of RepeatMasker which filteres highly repetitive sequences; [@] oligo similarity of the second target is less than 40%; * GC content constrains 28 ± 6%.

2.6.1 Design of gene and exon specific DNA microarrays for *Plasmodium vivax*

Plasmodium vivax causes debilitating disease that impairs the quality of life and

economic productivity of large regions of the South East Asia and South America (Price, et al., 2007). In many of these regions *P. vivax* is the most prevalent species of malaria. *P. vivax* has a similar life-cycle to the more fatal species of human malaria *P. falciparum*. However, several notable differences exist between these two species. These include a preference of *P. vivax* for reticulocytes, and the presence of persistent liver forms, hypnozoites, which can cause relapse weeks after an initial infection (Mendis, et al., 2001). The intrachromosomal regions of the *P. vivax* genome has a significantly higher GC content than *P. falciparum* with approximately 55% of AT, however, the subtelomeric region are comparable to the *P. falciparum* genome with approximately 80% AT content (Carlton, 2003). These extreme GC content fluctuations create major obstacle for the design of a balance of oligonucleotides with even distribution of Tm that are necessary for microarray assemblies. Here we use the OligoRankPick program to design gene specific long oligonucleotide probes (60 nt) with defined 40% GC-content for the entire genome. Total 9810 oligos were selected for 5341 genes, in which 9746 are unique (99.3%). Based on both criteria (<= 40% similarity and 5% deviation of GC content), 9309 oligonucleotides (95%) representing 5167 genes (97.6% of the genome) matched the desired parameters. Using this design, the complete transcriptional profile throughout the intraerythrocytic developmental cycle (9 time-points, IDC) of *P. vivax* was analyzed. The transcriptional regulation cascade of the *P. vivax* synthenic genes resembles the previously reported *P. falciparum* IDC transcriptome in which each cellular function is timed to a specific developmental

stage. In contrast the global distribution of mRNA abundance of the non-syntenic genes exhibits a strong bias towards the extremes of the IDC; the schizont – ring stage transition (Bozdech, et al., 2008).

2.6.2 Design of a cross-species gene specific DNA microarrays for rodent malaria parasites *P. yoelii, P. chabaudi and P. berghei*

In the past decade basic biological knowledge of the rodent models of malaria lags behind that of *P. falciparum*. The completions of the genomes of the three rodent malaria species (*P. berghei*, *P. chabaudi* and *P. yoelii*) increased the interest in these model organisms pointing out many similarities that can be exploited for biological research on human malaria. The discovered high level of homology as well as synteny between the *Plasmodium* species opened the door to many in functional genomics projects, such as comparative genomics and transcriptomic analyses. We are devoted to design DNA oligonucleotides to represent three genomes on one chip. This cross-species gene specific DNA microarray would facilitate the comparative research of the syntenic genes between the three rodent malaria genomes as well as the non-syntenic or species-specific genes.

Figure 2.7 summarizes the global overview on the design. Essentially, the three rodent malaria parasite genomes were assembled using the PHRAP program (http://www.phrap.com). Subsequently, OligoRankPick (Hu, et al., 2007) was used to design oligonucleotide probes for three entire genomes. This was shown to have significant improvements over other algorithms in oligonucleotide design even

when dealing with large fluctuations in GC content and abundant gene duplications. There are about 7861 predicted coding sequences in *P. yoelii* genome, 12216 in *P. berghei* and 15095 in *P. chabaudi*. We first used all possible oligonucleotides of *P. yoelii* to search in the homologous region of the other two species using the NCBI-BLAST program (Altschul, et al., 1997) (figure 2.7A). The using four parameters: BLAST score (first and second match), GC content and self-annealing score to search the oligonucleotides (figure 2.7B). Each score is then transformed into a rank and a weighted rank-sum is calculated for each oligonucleotide with the final oligonucleotide being selected based on the smallest rank-sum value. These oligonucleotides were then used to select for those that are optimal for all three species, followed by oligos for *P. yoelii* and *P. berghei* and then for *P. yoelii* and *P. chabaudi*. Next, only oligos specific for *P. yoelii* were selected (figure 2.7A). These oligonucleotides were then used to mask all the *P. berghei* and *P. chabaudi* predicted coding sequences and the remaining sequences were used to design *P. berghei*-specific or *P. chabaudi*-specific oligonucleotides. A total of 17650 oligonucleotides were obtained and the breakdown is shown in Figure 2.8. In this oligonucleotide set, 5461 oligonucleotides can detect genes from three rodent malaria species of *P. yoelii*, *P. berghei* and *P. chabaudi* at the same time. Expression and CGH analysis of pan-rodent malaria based on this chip was presented in the Liew et al.'s paper (Liew et al. 2008 manuscript in preparation).

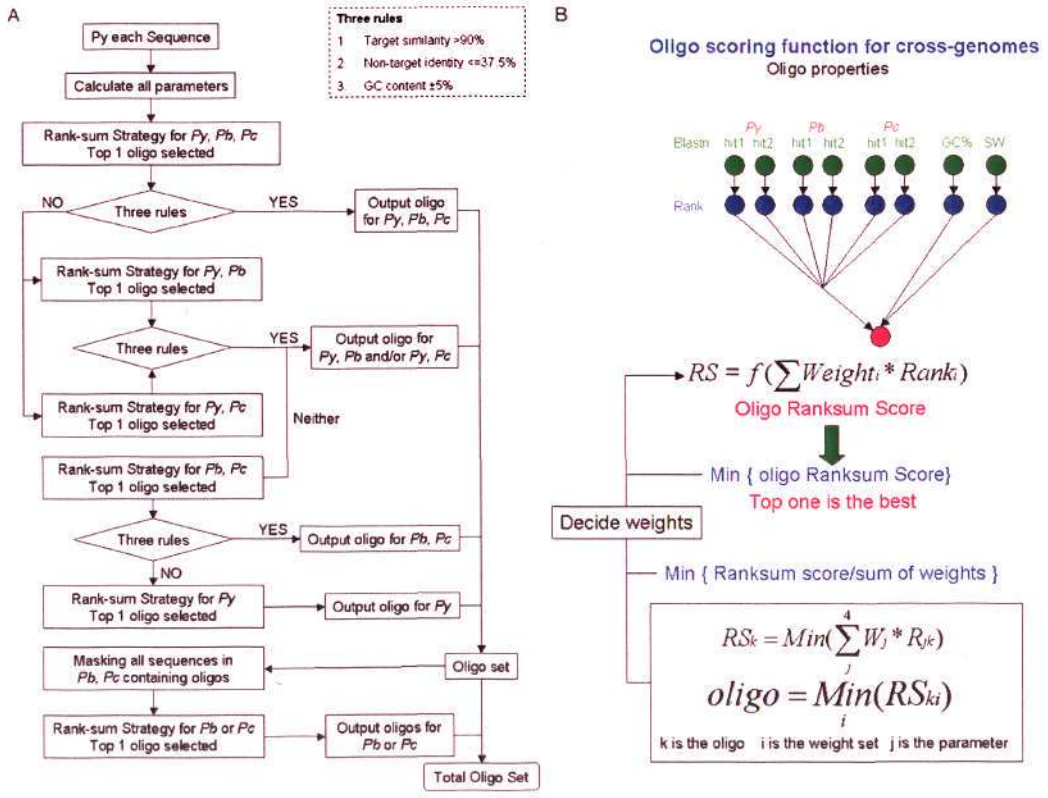


Figure 2.7 The overall design schematics of the pan-rodent chip. (A) Methodology of the chip design. First, all possible oligonucleotides of *P. yoelii* were used to search in the homologous region of the other two species using NCBI-BLAST and were scored and ranked accordingly. The oligonucleotides were then filtered using three rules: (i) at least 90% homology to target sequences, (ii) less than 37.5% to non-target sequences and (iii) GC% tolerance of ±5%. Oligonucleotides for all three species were selected followed by oligonucleotides for *P. yoelii* and *P. berghei* and then for *P. yoelii* and *P. chabaudi*. Next, the remaining oligonucleotides were selected to be specific only for *P. yoelii*. The remaining sequences unaccounted for were then used to design oligonucleotides specific to *P. berghei* or *P. chabaudi*. (B) Rank-sum strategy from OligoRankPick (Hu, et al., 2007).

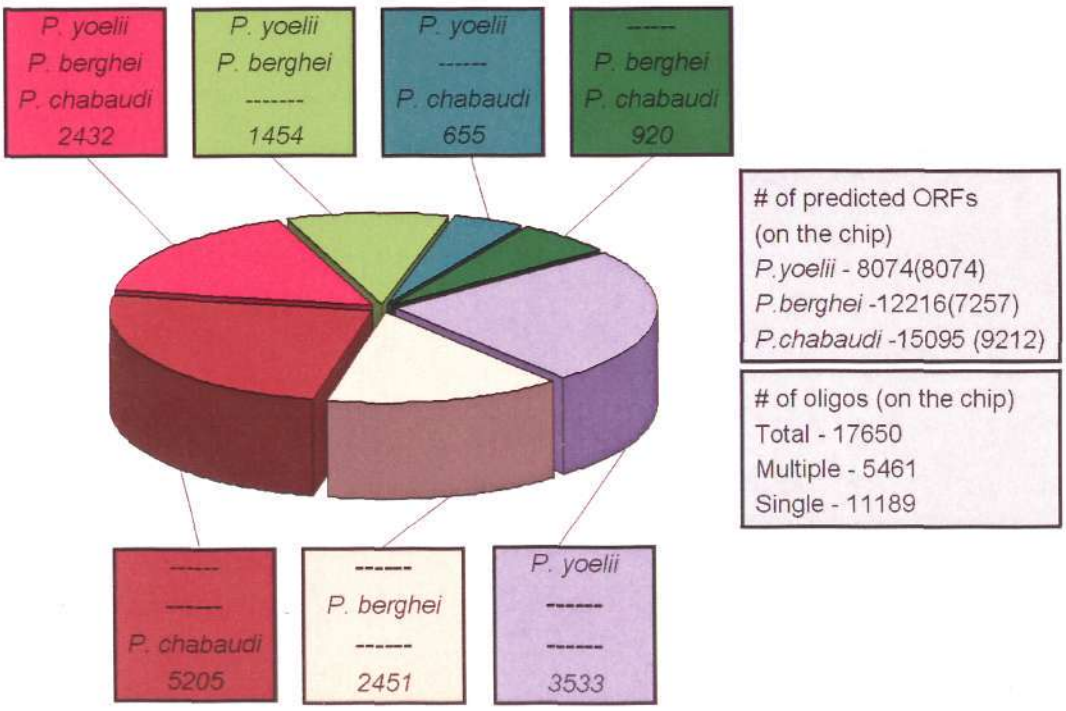


Figure 2.8. Statistical information of all oligonucleotides on the chip of rodent malaria. All oligonucleotides are 60 bases long and the GC content is fixed at 30% and the allowable deviation is 5% for overlapping oligonucleotides. Complementary oligonucleotides to each rodent malaria parasite species was calculated from the sum of all possible combinations, i.e. oligonucleotides specific to itself and those that can hybridize to itself and to other rodent malaria parasite species.

2.7 Conclusions and outlook

OligoRankPick provides a powerful alternative for long oligonucleotide microarray design for genomes with extreme GC content fluctuations and high abundance of highly homologous gene families. In its simplest implementation a user needs only to define the probe length and an expected GC content or Tm.

However, for specialized applications, OligoRankPick provides the user with the option of setting the range of relative importance (weight) of each parameter as well as optimization of the quality control target values. Using this method we have designed several long oligonucleotide DNA microarrays for the parasitic species including *P. falciparum, P. vivax* and pan-rodent malaria. Transcriptome analyses of two *P. falciparum* developmental stages demonstrated that the designed microarray provides the most comprehensive coverage of the *P. falciparum* genome presently available.

Although the actual oligonucleotide selection algorithm is highly efficient, the initial BLAST searches consume significantly high amounts of computer time. In our implementation, we isolate the time-consuming BLAST step (script oligoblast.pl), which can be run on different computers or a computer cluster. In the future, we hope to perform the BLAST searches with parallel processing methods such as mpiBLAST (http://mpiblast.lanl.gov/), which is much faster and more efficient to design oligonucleotides for large genomes like human genome. Another consideration is the incorporation of more novel models to evaluate the qualities of oligonucleotides. For example, hybridization energy model, Gibbs free energy model of DNA secondary structure. In the future, OligoRankPick would provide an interface for users to select and define these models.

2.8 Materials and methods

2.8.1 Genome sequences and annotations

The *E. coli* gene sequence file with 4237 CDSs and genomic sequence file were downloaded from the NCBI genome database. The *S. cerevisiae* gene sequence file with 6680 CDSs, and whole genome sequence file were downloaded from the ENSEMBL database (www.ensembl.org). The protein coding sequence files and their whole genomic sequence files of *P. falciparum*, *P. yoelii*, *P. berghei* and *P. chabaudi* were downloaded from PlasmoDB version 4.4 (www.plasmodb.org). The coding sequences and genomic sequences of P.vivax were downloaded from Tiger genome database (www.tigr.org) under the permission.

2.8.2 Microarray manufacture and hybridization

Microarray manufacture and hybridizations were conducted as previously described (Bozdech, et al., 2003). Briefly, all oligonucleotides in 384 well plates were printed on the polylysine-coded glass slides using BioRad microarray printer system. Printed slides were post-processed by rehydration, UV cross-linking and succinic anhydride (ALDRICH, Cat. 239690) block. The labeled cDNA samples were hybridized to the chip in MAUI system (BioMicro, Utah, United States) for 12-14 hours at 65°C. Data were acquired and analyzed by GenePix (Axon Instruments, Union City, California, United States). Array data were stored and normalized in Acurity 4.0 system (Axon Instruments, Union City, California, United States).

2.8.3 Quantitative real-time PCR

Real time RT PCR was performed in a total reaction volume of 20µl which contained 1µl cDNA template (10ng/µl), 0.5 µl forward and reverse primer (10µM), and 10µl of 2 x Power SYBR Green PCR Master Mix (Applied Biosystems). The temple cDNA was generated using the first strand cDNA synthesis protocol used for the microarray hybridization. For the amplification the universal thermal cycling parameters were programmed as follows: 5 min activation at 95°C, followed by 40 cycles of 20s at 95°C, 30s at 50°C, 40s at 72 °C and 1 min at 60°C. Each reaction was run in triplicates. The mRNA abundance ratios were calculated using ABI 7500 Fast Real-Time PCR Systems and the relative quantitation of gene expression was performed using the comparative CT method. Primers for PCR were designed using DNAMAN (Lynnon Corporation).

Chapter 3 Transcriptional profiling of growth perturbations and a functional interactome network of human malaria parasites, *Plasmodium falciparum*

3.1 Summary

More than 50% of the genes of *Plasmodium falciparum*, the deadliest form of human malaria, are of unknown function. To create a function interactome network, we analysed the global transcription response of *Plasmodium* cells to 20 compounds in 29 independent time series, creating 183 microarray data points. We demonstrate that at least half of the *Plasmodium* genome can respond to at least one of these growth perturbations and that functionally related genes share similar transcriptional profiles. To reconstruct a high-confidence probabilistic interactome network we integrated the transcription data with phylogenetic profiles, domain interaction linkages and the yeast two-hybrid results. Using this network we predict the function of 2545 *Plasmodium* hypothetical proteins. To validate our network we retrieved 263 new proteins linked with merozoite invasion, a process which is considered as a key target for malaria control. Intracellular localization of a subset of these proteins confirms their function in this process.

3.2 Introduction

A fundamental problem in systems biology is to infer biological functions for the high number of uncharacterized proteins that are identified by large scale genome

sequencing but lack significant sequence or structural homologies to other known proteins. In pathogenic organisms, understanding of molecular and biochemical processes is of particular importance for the design and development of new drugs and vaccine based intervention strategies (Murali, et al., 2006). Due to this, several computational techniques for functional gene predictions that utilize data from genome-wide experimental approaches were developed in recent years (Enright, et al., 1999; Ge, et al., 2001; Kemmeren, et al., 2002; Marcotte, et al., 1999; Ponting and Dickens, 2001; Sharan, et al., 2007; Valencia and Pazos, 2002). These techniques integrate data from functional genomics (e. g. transcriptional profiling, two hybrid-screens) or proteomics (e .g. shot gun mass spectroscopy surveys) and multiple types of bioinformatics studies (e. g. domain predictions, phylogenetic profiles). The main purpose is to generate functional-linkage networks between proteins/genes in order to predict biological relevance for the genes whose sequence does not provide any direct functional clues. The most evolved approaches serve as a powerful reference for functional prediction of uncharacterized genes based on their position in the network by evaluating their proximal functionalities (Chua, et al., 2006; Karaoz, et al., 2004; Schwikowski, et al., 2000; Sharan, et al., 2007). These network approaches were also shown to significantly contribute to the understanding of biological mechanisms associated with disease processes (Calvano, et al., 2005; Pujana, et al., 2007), and for investigating how a cell adapts to changing environments (Guo, et al., 2007; Whitehead, et al., 2006).

In the causative organism of the deadliest form of human malaria, *Plasmodium falciparum*, more than 50% of all genes are still functionally uncharacterized due to their lack of sequence homology with known genes in other organisms (Gardner, et al., 2002). Recognizing this deficit, Date and Stoeckert (Date and Stoeckert, 2006) constructed a first interactome network involving the transcriptome data of the *P. falciparum* intraerythrocytic developmental cycle (IDC) (Bozdech, et al., 2003) and genomic context data that included phylogenetic profiles and Rosetta stone data. Using a naïve Bayesian method, it was possible to reconstruct a functional network including 3667 proteins at the 50% confidence level (Date and Stoeckert, 2006). In a following study, a probabilistic gene interaction network (interactome) was assembled incorporating evolutionarily conserved protein linkages, derived from *in silico* domain-domain interaction predictions and experimental protein-protein interactions based on the yeast two hybrid system survey (Wuchty and Ipsaro, 2007). In this case, the assembled interactome included 2321 proteins which accounts for approximately half of the *P. falciparum* genome. Although these two networks provided a significant contribution to gene annotation, several limitations severely hampered their impact. The main caveat of the Date and Stockert's interactome is the nature of the transcriptional data which postulate high correlations for many functionally un-related genes due to the monotonous character of transcriptional regulation during the *P. falciparum* IDC. The calculation of a regulatory network based on the IDC transcriptome with an average connectivity of 30 resulted in a higher value of a Pearson correlation

coefficient (PCC) threshold of 0.95, which is unreasonably high compared to other organisms (Khanin and Wit, 2007). The second interactome contains the small number of genes for which appropriate information exists and a high number of false positive results in the yeast two hybrid system data (Wuchty and Ipsaro, 2007).

Microarray analyses of global transcriptional responses to growth perturbations provide a powerful input that can significantly improve the accuracy and proteome coverage of probabilistic interactome networks (Hughes, et al., 2000; MacCarthy, et al., 2005; Zak, et al., 2003). Given the limitations of the life cycle based transcriptome data, growth perturbation data were suggested to be extremely helpful for *Plasmodium* systems biology approaches (Winzeler, 2006). Until today, very little is known about transcriptional responses of *P. falciparum* to growth perturbations. It was shown that exposure of *P. falciparum* cells to chloroquine induces only low amplitude transcriptional changes of a wide spectrum of functionally unrelated genes (Gunasekera, et al., 2007). A similar lack of a specific signature response was also observed in *P. falciparum* cells exposed to a protein kinase inhibitor that is otherwise capable of inhibiting parasite growth (Kato, et al., 2008). In contrast to these studies, Oakley *et al* demonstrated that febrile temperatures induce a more specific transcriptional response that involves 336 *P. falciparum* genes including genes encoding membrane-associated proteins that are exported to the host cell cytoplasm and likely affect parasite sequestration and antigenic presentation (Oakley, et al., 2007). In addition, these transcriptional

responses included factors of protein stability and trafficking, RNA metabolism, signal transduction, nuclear functions and general metabolism. Using these data Oakley *et al* was able to predict and partially validate putative functions for ~100 previously uncharacterized proteins (Oakley, et al., 2007). These studies demonstrated the potential of growth perturbation analyses in *Plasmodium* cells for gene annotation purposes.

The main rationale of this study was to assemble an interactome network of *P. falciparum* gene/proteins that incorporate data from extensive transcriptional profiling of growth perturbations using a wide array of small molecular inhibitors. Here we used 20 diverse small molecular compounds to inhibit the growth and/or development of *P. falciparum* during the asexual erythrocytic developmental stages. These included inhibitors of enzymatic activities (e.g. proteases and protein kinases and histone deacetylases), general cellular functions (e.g. microtubule and membrane formation and intracellular Calcium concentration) and several common antimalarial drugs (for complete experimental panel see Table S1). Combining the transcriptional profile dataset with *in silico* generated phylogenetic profiles, domain-domain interaction evidence and the yeast two-hybrid system-based protein-protein interactions, we constructed a high confidence gene interactome network using a probabilistic Bayesian network approach. Based on this network, we assigned function to 2545 hypothetical proteins using a weighted neighbor counting method. Using life cell imaging, we were able to verify functional assignments of 19 (out of 21) proteins that were predicted to be localized in the

cellular compartments associated with the *Plasmodium* invasion machinery.

3.3 Results

3.3.1 Transcriptional profiling of growth perturbations

In the first step we carried out microarray measurements of global transcriptional responses of *P. falciparum* to twenty growth-inhibitory compounds. These included the common anti-malarial drugs: chloroquine (CQ), quinine (Q), artemisinine (ART) and the experimental compound febrifugine (FEB); and small molecular inhibitors that inhibit parasite growth and/or development: protease inhibitors (E64, leupeptine and PMSF), protein kinase inhibitors (ML7, W7), histone deacetylase (HDAC) inhibitors (Apicidin and TrichostatinA) and inhibitors of cation-dependent APTases (Na_3VO_4), microtubule (colchicine), and membrane formation (Retinol A) (table 3.1). For each compound, we carried out a treatment time course where synchronized *Plasmodium* cells were exposed to IC50 or IC90 concentrations (with two exceptions, table 3.1). The IC50 and IC90 concentrations were determined individually for each compound and the culturing conditions used during the transcriptional profiling (2% hematocrit with 5% parasitemia). The final dataset included 29 time courses analyzed by 183 individual microarray measurements (figure 3.1A).

Across the entire experimental panel, 3226 genes exhibited at least 3-fold change in the transcript level in at least one time point of one growth perturbation (figure 3.1A). The first striking feature of these results is the large diversity of the global

transcriptional changes induced by the different compounds. Several compounds stimulated high amplitude transcriptional responses which involve narrow but well defined groups of genes. The most striking examples are FK506 and Cyclosporine A which induce 256 and 189 and suppress 29 and 42 genes by >3-fold, respectively (figure 3.1A). The substantial overlap between the gene groups induced by these two compounds is consistent with the presumed mode of action of both inhibitors; suppression of the Calcineurin-dependent signaling pathway (Bell, et al., 2006; Kumar, et al., 2005; Liu, et al., 1991). Significant similarities were also observed between *P. falciparum* transcriptional responses to three inhibitors of calcium dependent signaling (ML7, W7, and KN93) (figure 3.1A). This indicated that similar to Calcineurin, calcium/calmodulin-dependent protein kinases (CDPK) are linked with transcriptional regulation of the parasite. Interestingly, there was only a limited overlap between the transcriptional responses induced by the CDPK, and Calcineurin inhibitors. This suggested that these two types of intracellular signaling pathways play specific, largely non-overlapping roles in *Plasmodium* parasites. Although both classes of inhibitors caused an arrest of schizont rupture (figure 4.1), none of the transcriptional changes induced by these inhibitors were consistent with a general arrest of the IDC transcriptional cascade (figure S3.1). This apparent contradiction could be explained by the time difference between the two observations. While the transcriptional analyses were carried out during the first 8 hours post treatments, starting 32-33 hours post invasion (hpi), the arrest of morphological development was observed only 14 hr

Table 3.1 Summary of microarray data sets used in the analysis

Exp.*	Category	#Exp.	Decription
	Control	7	Dd2, start time-point 32hpi, time course 1,2,4,6,8,10,12pt
1	ML-7	7	Dd2, IC50 ~ 1.2 uM
2	W-7	7	Dd2, IC50 ~ 1.2 uM
3	KN93	7	Dd2, IC50 ~ 1.2 uM
4	Staurosporine	7	Dd2, IC50 ~ 80 nM
	Control	8	Dd2, start time-point 33hpi, time course 1,2,4,6,8,10,12,14pt
5	Cyclosporine A	8	Dd2, IC50 ~ 88 nM
6	FK506	8	Dd2, IC50 ~ 118 nM
7	Roscovitine A	8	Dd2, IC50 ~ 1.6 uM
	Control	6	Dd2, start time-point 18hpi, time course 1,2,4,6,8,10pt
8	Chloroquine	6	Dd2, IC50 ~ 43nM
9	Quinine	6	Dd2, IC50 ~ 44nM
10	Febrifugine	6	Dd2, IC50 ~ 4.5nM
11	Artimesinin	6	Dd2, IC50 ~ 28M
	Chloroquine	6	3D7, start time-point 18hpi, time course 1,2,3,4,6,8pt
12		6	3D7, IC50 ~ 41nM
13		6	3D7, IC90 ~ 72nM
14		6	3D7, 2*IC90 ~144nM
	EGTA	10	Dd2, start time-point 34hpi, time course 1,2,3,4,5,6,7,8,9,10pt
15		10	Dd2, IC50 ~ 0.5mM
16		10	Dd2, IC90 ~ 3.5mM
	Trichostatin A	6	Dd2, start time-point 34hpi, time course 1,2,3,4,6,8pt
17		6	Dd2, IC50 ~ 25nM
18		6	Dd2, IC90 ~ 51nM
	Apicidin	6	Dd2, start time-point 18hpi, time course 1,2,3,4,5,6pt
19		6	Dd2, IC50 ~ 20nM
20		6	Dd2, IC90 ~ 70nM
	Apicidin	5	Dd2, start time-point 34hpi, time course 1,2,3,4,5pt
21		5	Dd2, IC50 ~ 23nM
22		5	Dd2, IC90 ~ 85nM
	control	5	Dd2, start time-point 33hpi, time course 0.5,1,2,3,4pt
23	E64	5	Dd2, IC50 ~ 3.2uM
24	PMSF	5	Dd2, IC50 ~ 1.32mM
25	Leupeptine	5	Dd2, IC50 ~ 5.4uM
26	Retinol A	5	Dd2, IC50 ~ 107uM
	control	5	Dd2, start time-point 33hpi, time course 0.5,1,2,3,4pt
27	Colchicine	5	Dd2, IC50 ~ 46.3uM
28	Na3VO4	5	Dd2, IC50 ~ 17uM
29	Staurosporine	5	Dd2, IC50 ~ 87uM
	Field strains	24	cell cycle of field strains from Africa
	Lab strains	18	Life cycle of lab strains (3D7/Dd2/T996) Time-points of very 8 hour
	Life cycle*	148	IDC transcriptome of lab strains (3D7/Dd2/ HB3), time-points of very 1 hour
	Total	437	

* Sachel Mok performed Exp.8-14; Sabna Cheemadan performed Exp.15-16; Brigitta performed Exp.17-22; Balbir Chaal performed Exp.23-29.

after the addition to the drug during the schizont rupture (figure 4.1). Thus the observed transcriptional changes might represent the initial specific transcriptional response of *P. falciparum* parasites that is subsequently followed by a developmental arrest and cell death. This is in a sharp contrast with the EGTA treatment which also caused a rupture arrest but essentially no transcriptional changes during the early part of the treatment (figure 3.1A and data not shown). Although further studies are required to understand the role of CDPK and Calcineurin signaling pathways in the progression of the *P. falciparum* life cycle, these data suggest their importance for this process.

In contrast to the protein kinase inhibitors, a number of compounds had only a subtle effect on gene expression despite their strong growth inhibitory properties in *P. falciparum* parasites. These included colchicine, Na_3VO_4, E64 and Leupeptine (figure 3.1A). Similarly to these compounds, all three malaria drugs, chloroquine, quinine and artemisinine induced only low amplitude transcriptional responses that involved relatively small numbers of genes (figure 3.1A). In agreement with previous analyses by Gunasekera and colleagues, the transcriptional responses to chloroquine were highly reproducible and dose dependent (Gunasekera, et al., 2007). A total of 26, 49, and 87 genes were induced by >3-fold (194, 257 and 330 genes by >2-fold) with IC50, IC90, and 2*IC90 concentrations of chloroquine, respectively. These observations suggest that at least a portion of these low amplitude transcriptional changes might reflect relevant physiological response to the drug. The minor effect of E64 and Leupeptine on *Plasmodium* transcription is

surprising since PMSF, a generic protease inhibitor, induced a broad transcriptional response that is consistent with an arrest of the IDC transcriptional cascade (figure 3.1A and S3.1). These data indicate that in contrast to E64 and Leupeptine, PMSF has an additional target that is potentially linked to the regulatory pathways of malaria parasites.

Out of the 20 inhibitors, 4 treatments (EGTA and PMSF, staurosporine and TrichostatinA) caused IDC developmental arrest (figure S3.1). EGTA that is thought to deplete intra and extracellular Ca^{2+} (100mM ~ IC50) was found to block parasite egress from mature schizonts and at the same time to retain the entire transcriptional profile of the late schizont stage even 6 hours after the estimated time of rupture and reinvasion (figure 3.1A). In addition to the IDC arrest, treatments of *P. falciparum* cells with apicidine (HDAC inhibitor) caused a general deregulation of the IDC transcriptional cascade by de-repression of genes that are normally suppressed at both studied developmental stages (trophozoite and schizont) (figure 3.1A). Interestingly, TrichostatinA, another HDAC inhibitor, induced broad transcriptional changes that consist of both IDC arrest (figure S3.1) and deregulation (figure 3.1A). The effect of the HDAC inhibitors on chromatin remodeling and transcriptional regulation in *P. falciparum* is presently under investigation in our laboratory (Chaal et al manuscript in preparation).

Taken together, the growth perturbation analyses illustrate a complex character of *Plasmodium* responses to environmental perturbations. On the one hand, there are a number of cellular functionalities (such as protein phosphorylation

pathway) firmly linked to the *Plasmodium* transcriptional regulation and their inhibitions induce vigorous responses involving a substantial amount of genes. On the other hand, disruptions of other *Plasmodium* biological functions (such as microtubule formation, ATP hydrolysis or a subgroup of proteases) produce only a limited low amplitude reaction by the *P. falciparum* transcriptional machinery. There are also several functionalities whose disruption is associated with the global transcriptional regulation of the *P. falciparum* life cycle either via a developmental arrest (EGTA) or via perturbation of histone modifications (HDAC inhibitors).

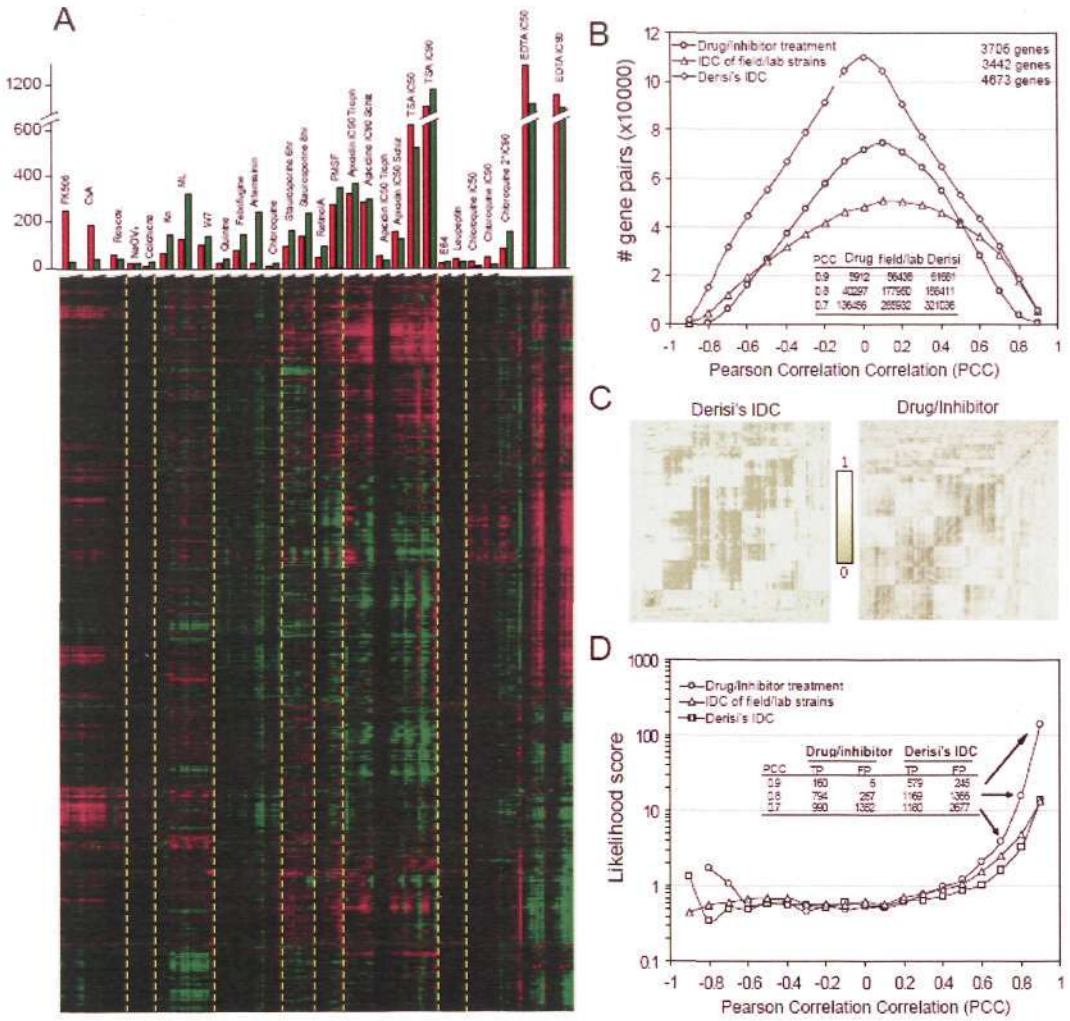


Figure 3.1. Overview of the gene expression responses of *P. falciparum* to growth

perturbation induced by drug or inhibitor treatments. A. The heatmap summarizes results from 184 microarrays from 29 time courses monitoring transcriptional changes in *P. falciparum* induced by 20 small molecular inhibitors (experimental summary see table 3.1). The treatment experiments were conducted in the time courses (ordered along the horizontal axis) and genes were arranged using hierarchical clustering. A total of 3226 genes which show at least a 3-fold change of mRNA abundance in at least one experiment are included in the overview dataset. The bar diagram (top) indicates the total number of genes that show >3-fold up-regulation (red bar) or down-regulation (green bar) for each treatment experiment. The treatment experiments were ordered and grouped (yellow dashed lines) according the similarity of the transcriptional response. B. The Pearson Correlation Coefficient (PCC) distributions of gene expression profiles from the Drug/inhibitor treatments, the publicly available IDC transcriptome (Derisi's IDC) of 3D7, Dd2 and HB3 (Llinas, et al., 2006), and additional IDC transcriptomes for three field and three laboratory strains generated in our laboratory. The number of gene pairs in the PCC bins >0.7 is indicated in the inset table. C. Heat map of hierarchical clustering of the PCC profiles. A PCC profile was assembled for each gene as a function of correlations of its expression profile with every other expression profile in the dataset (perturbation or IDC datasets). The PCC profiles were subjected to hierarchical clustering that reveal natural grouping of highly correlated genes (small distance ~ dark green) and distinguish from uncorrelated gene groups (large distance ~ white). D. Co-expressed genes in drug or inhibitor microarrays had higher likelihood scores than in IDC transcriptomes. Likelihood score was used as a function of ratio of observed positive probability to negative probability based on the functional KEGG data to measure the functional association for the co-expressed genes at a different level (different PPC).The number of false positive (FP) and true positive (TP) predictions of gene linkages are indicated in the inset table.

3.3.2 Co-expression of functionally related genes across the perturbation panel

To evaluate the complexity of the growth perturbation transcriptional data we calculated Pearson Correlation Coefficients (PCC) between the expression profiles across the entire experiment for each gene pair in the perturbation dataset as well as in the high resolution (1hour) DeRisi IDC transcriptome and low resolution (8hour) IDC transcriptomes of three laboratory strains (3D7, T996, Dd2) and three short term culture adapted field isolates. Compared to the IDC transcriptome, we observed a tighter peak of the distribution of the PCC values from the perturbation data compared to the IDC transcriptomes. In particular, there is approximately a 3 and 4.5 -fold decrease in the number of gene pairs with the PCC intervals <0.7-0.8> and <0.8-0.9>, respectively, and more than 10-fold decrease in the number of gene pairs with PCC $\geq$ 0.9 in the perturbation dataset compared to the IDC transcriptomes (figure 3.1B). A similar drop in the gene pair number was observed in the negative correlation PCC bins (PCC $\leq$-0.7, data not shown). Moreover, hierarchical clustering of the PCC profiles for each gene pair in the dataset show a considerably tighter pattern in the IDC dataset compared to the perturbation data for which the pattern of the pair-wise distances is more dispersed (figure 3.2C). Taken together, this indicates that compared to the IDC transcriptome, the perturbation dataset defines narrower gene groups that share transcriptional regulation across the wide spectrum of growth conditions/perturbations. To systematically evaluate functional relationships of the transcriptionally co-regulated genes we utilized the subset of genes with a

functional prediction defined for 492 genes in 71 pathways defined by the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa, et al., 2004) (see materials and methods). For these we calculated a likelihood score as a function of a ratio between the number of positive (gene pairs in common pathways) and the number of negative observations (gene pairs not in common pathways) for different PCC thresholds (figure 3.1D). The results show that transcriptionally co-regulated genes exhibit a significantly lower rate of false positives and thus higher likelihood scores compared to the IDC transcriptomes. Overall there are 1.2, 1.5, and 3.6-times less gene pairs in the 0.7, 0.8, 0.9 PCC bins of the perturbation dataset compared to the IDC datasets. Amongst these, the false positive rate was improved by 1.6, 3.5 and 11 -fold (figure 3.1D). These data suggest that the transcriptional profiling of the chemical perturbations of *P. falciparum* growth have a high predictive accuracy of functionally related genes based on their transcriptional regulation.

3.3.3 Reconstruction of a probabilistic gene functional network

To fully utilize the potential of the perturbation transcriptional profiling for functional gene predictions we assembled a probabilistic network which integrates these results with three additional datasets. The first dataset represented phylogenetic profiles which consist of sequence homology values (E-values) of all 5363 *P. falciparum* protein sequences to their orthologues in 210 completely sequenced, publicly available genomes. Using the mutual information method

(Date and Marcotte, 2003; Sun, et al., 2005) we identified 12,406,623 phylogenetic profiles for 4983 proteins (figure S3.2). The second dataset included evidence of domain-domain interactions predicted in the deduced amino acid sequence of all *P. falciparum* proteins by HMM based searches within the PFAM database (Sonnhammer, et al., 1998) and the Lee's domain-domain interaction dataset (Lee, et al., 2006). A total of 179,481 linkages between *P. falciparum* proteins were defined by this approach (data not shown). The third dataset includes the experimental observations of the 2811 protein interactions for 1308 proteins in *P. falciparum* that were detected by yeast two-hybrid system screens (LaCount, et al., 2005). In addition to these datasets, the perturbation microarray data was combined with the publically available IDC transcriptomes from three *P. falciparum* strains, 3D7, HB3 and Dd2 (Llinas, et al., 2006). For this, PCCs between the IDC expression profiles were merged with the PCCs from the perturbation dataset using an optional average approach (materials and methods).

The potential of forming a protein-protein functional interaction was scored for each individual dataset as the probability of each linkage to fall into the positive or the negative benchmark (figure S3.3 and table S3.1). The final likelihood score for each protein linkage was generated by integrating the four likelihood scores using a Bayesian method (Jansen, et al., 2003; Lee, et al., 2004). Overall likelihood scores for 14,168,597 functional linkages between 5374 *P. falciparum* proteins (99.2 % of the proteome) were calculated based on evidence captured by at least one of the derived datasets. Information for 4774 proteins (88.1%) was represented in the

perturbation transcriptional profiling and at least one other dataset while 4785 proteins (88.3%) were represented in any two individual data sets. In general, the integrated likelihood scores provide higher proteome coverage than each of the individual input datasets at all probability thresholds (figure 3.2A).

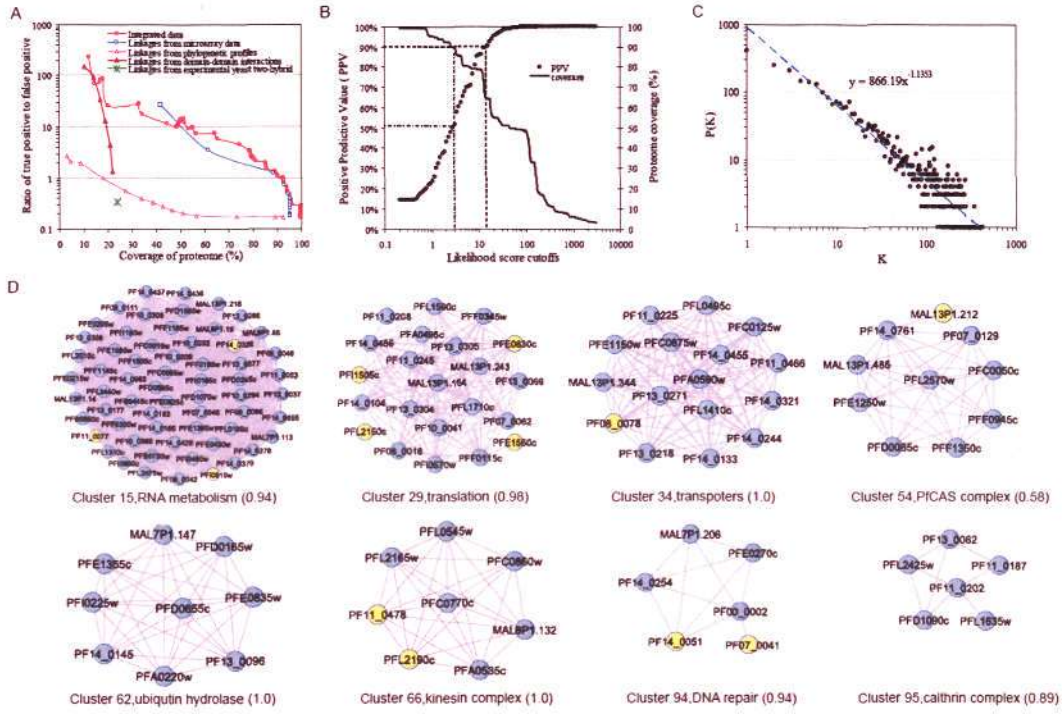


Figure 3.2 Reconstruction of the PlasmoINT interactome network. A. For each data type, we calculate proteome covered as a function of the ratio between the observed true and false positive observations when compared against benchmarks dataset (492 gene assigned to 71 KEGG pathways, see materials and methods). Integration of different functional datasets leads to higher genome coverage and accuracy than any of the individual methods. B. The predictive precision rates (Predictive Positive Value, PPV) at different likelihood score cutoffs were evaluated by 10-fold cross validation and the proteome coverage of the integrated functional dataset. PPVs were plotted as a function of the likelihood score cutoffs. PPV is calculated as the ratio of observed true positive number (TP) to the total number of TP and false positives (FP). Each dot of the ratios represents an average of ten cross-validations at a particular likelihood score cutoff. The vertical dashed-line showed the likelihood score cutoffs and proteome coverages

corresponding to positive predictive value (PPV=TP/(TP+FP)) 50% and 90% (likelihood score thresholds of 3 and 14.5). At these ratios, true positive to false positive (TP to FP) was equal to 1 (~50% confidence) and 9 (~90% confidence), respectively. C. Characterizations of the network topological structures in the 90% confidence network (for the 50% confidence network see Figure S5) expressed as the distribution of node connectivity. The scatter plot illustrates the gene numbers (P(K)) with the corresponding number of linkages (K). D. Examples of MCL modules identified in the 90% confidence network. In each modules, functionally characterized genes (purple circles) were examined in order to derive a coherence score representing the fraction of gene linkages that belong to a common functional groups e.g. RNA metabolism (coherence score 0.94) for cluster 15. The functionally uncharacterized (hypothetical) genes (yellow circles) provide suitable candidates for additional factors of a particular cellular or metabolic functionality.

For approximately 10% of proteins, the domain-domain interaction dataset generates high accuracy predictions. However, the proteome coverage of this dataset is limited due to the fact that only a small fraction of *P. falciparum* proteins contained well conserved functional domains (data not shown). In contrast, the transcriptome data and phylogenetic profiles can provide high proteome coverage but their predictive values are consistently lower than the integrated likelihood scores (figure 3.2A). In our calculations, we observed low accuracy and low proteome coverage of the protein-protein interaction dataset based on the two-hybrid system that thereby provides a low contribution to the final likelihood scores (figure 3.2A and S3.3). In the 50% confidence network, only 299 linkages have a two-hybrid system component and omitting this dataset leads to a loss of

177 linkages (data not shown). Despite its low impact, we found it important to include this dataset as it might reveal novel functional linkages that could not be discovered otherwise.

Using the calculated functional linkages we assembled interactome networks based on two likelihood score thresholds for the 50% and 90% confidence precision rate (figure 3.2B). While the 50% precision rate predictions include 339,721 functional linkages between 4817 proteins (89% of the *P. falciparum* proteome), the 90% precision rate predictions define 72,748 linkages between 3475 genes (64%). The connectivity of both, the 50% and 90% confidence networks fits a power-law distribution with the power λ equal to 0.93 and 1.14, respectively (Figure 3.2C and S3.4). Its structure reflects a typical scale-free network without obvious hierarchical topological structure (figure S3.4), which suggests the existence of a relatively small number of highly connected nodes (hubs) in the *Plasmodium* gene functional network.

Table 3.2 summarizes comparisons between the newly assembled interactome network, termed PlasmoINT, and the previously assembled network by Date and Stoeckert termed PlasmoMAP (Date and Stoeckert, 2006). Overall, PlasmoINT provides a substantially improved proteome coverage as well as precision of linkage prediction. First, the 50% precision networks from both studies contain a comparable number of genes and linkages. However, PlasmoINT contains considerably more linkages that originate from two or more types of evidence (e.g. transcriptional profiling, phylogenetic profiling, domain prediction

and two hybrid system). Second, the 90% accuracy PlasmoINT network contains approximately 3 times more genes and 6 times more linkages compared to PlasmoMAP (table 3.2).

Table 3.2. Network comparison of PlasmoINT and PlasoMAP.

	PlasoMAP	PlasmoINT
Combined evidence	IDC transcriptome; Phylogenetic profiles; Gene fusion data	247 drug/inhibitor microarrays; Transcriptome of field strains; IDC transcriptome; Experimental PPI; Phylogenetic profiles; Domain-domain interactions
50% precision rate network		
Total linkages	388,969	339,721
At least two evidences	117,764 (30%)	309,670 (~91%)
Total proteins	3667 (~62%)	4817 (~89%)
90% precision rate network		
Total linkages	12,290	72,748
At least two evidences	12,034 (97%)	62,176 (~85%)
Total proteins	1415 (~26%)	3475 (~64%)
		Comparison of 50% precision networks
Genes present in both	3284 (90%)	3284 (68%)
Linkage present	341,224	188,798
Shared linkages	78,571 (23%)	78,571 (42%)
Lost linkages	262,253 (77%)	-
Gain linkages	-	110,227 (58%)
		Comparison of 90% precision networks
Genes present in both	1149 (81%)	1149(33%)
Linkage present	10,042	188,798
Shared linkages	2,303 (23%)	2,303 (12%)
Lost linkages	7,739 (77%)	-
Gain linkages	-	16,703 (88%)

Abbrev.: IDC, intraerythrocytic developmental life cycle; PPI, protein-protein interaction.

In both PlasmoMAP and PlasmoINT, the majority of the 90% accuracy linkages originate from at least two types of evidence, which illustrates the importance of the integration of the likelihood scores from multiple datasets in order to achieve

high confidence predictions. Third, direct comparisons between the two networks revealed that the majority of proteins found in PlasmoMAP were also represented in PlasmoINT. However, there were considerable differences in the structure of the linkages between the genes present in both networks with only a small fraction of PlasmoINT linkages replicated in PlasmoMAP (table 3.2). These data are consistent with our predictions that the fundamental improvements provided in PlasmoINT help to suppress the number of false positive results in the low (50%) precision network but boosts the overall number of linkages with high accuracy (90% precision rate). This observation is supported by the overall increase in the number of linkages in gene groups predicted in the functional annotations based on the Gene Onthology (GO), Malaria Parasite Metabolic Pathways (MPMP) (Ginsburg, 2008), KEGG databases (figure S3.5). Moreover, PlasmoINT provides much improved reconstruction of metabolic and cellular pathways by covering a large number of genes assigned to the functional groups (figure S3.5 and examples in S3.6).

3.3.4 Modular analysis and network-based gene function predictions

Accumulating evidence suggests that biological systems are composed of interacting modules that can group various cellular components into biologically relevant functional categories (Barabasi and Oltvai, 2004; Hartwell, et al., 1999). In the next step, we used an unsupervised graph clustering algorithm, Markov Cluster (MCL) algorithm (Brohee and van Helden, 2006; Enright, et al., 2002;

Krogan, et al., 2006) to identify 208 such modules in the 90% confidence network. In order to evaluate the biological relevance of the identified modules, we calculated a functional coherence (enrichment) score for 105 of the modules in which the functionally annotated genes are linked (figure 3.2D). This score represents the fraction of gene pairs that share functional annotations in a given module (for full list see figure S3.7). The top 35 modules with coherence score ≥ 0.4 (figure 3.3A) include many gene groups involved in basic metabolic processes such as redox and pyrimidine metabolism, ribosomal structure, DNA repair and classical clathrin-mediated vesicular transport. In addition, several *Plasmodium* specific functions such as rosette formation and mitochondrial and apicoplast membrane transport (figure 3.3A) can be deciphered by the MCL method in the 90% confidence network. These observations suggest that the assembled network detects functionally related genes with sufficient precision that it can be further explored for the functional annotation of functionally unidentified genes (figure 3.2D).

In the next step, we applied the Weighted Neighbor Counting (WNC) technique to the 50% predictive precision rate network in order to derive functional predictions of the 2662 hypothetical proteins present in this network. The choice of the 50% precision network instead of 90% was mainly driven by the higher proteome coverage of the low precision network. This approach takes advantage of the 2187 proteins with functional assignments based on 336 functional terms with more than 1 gene from the KEGG (70 terms), GO (145 terms for up to 7th

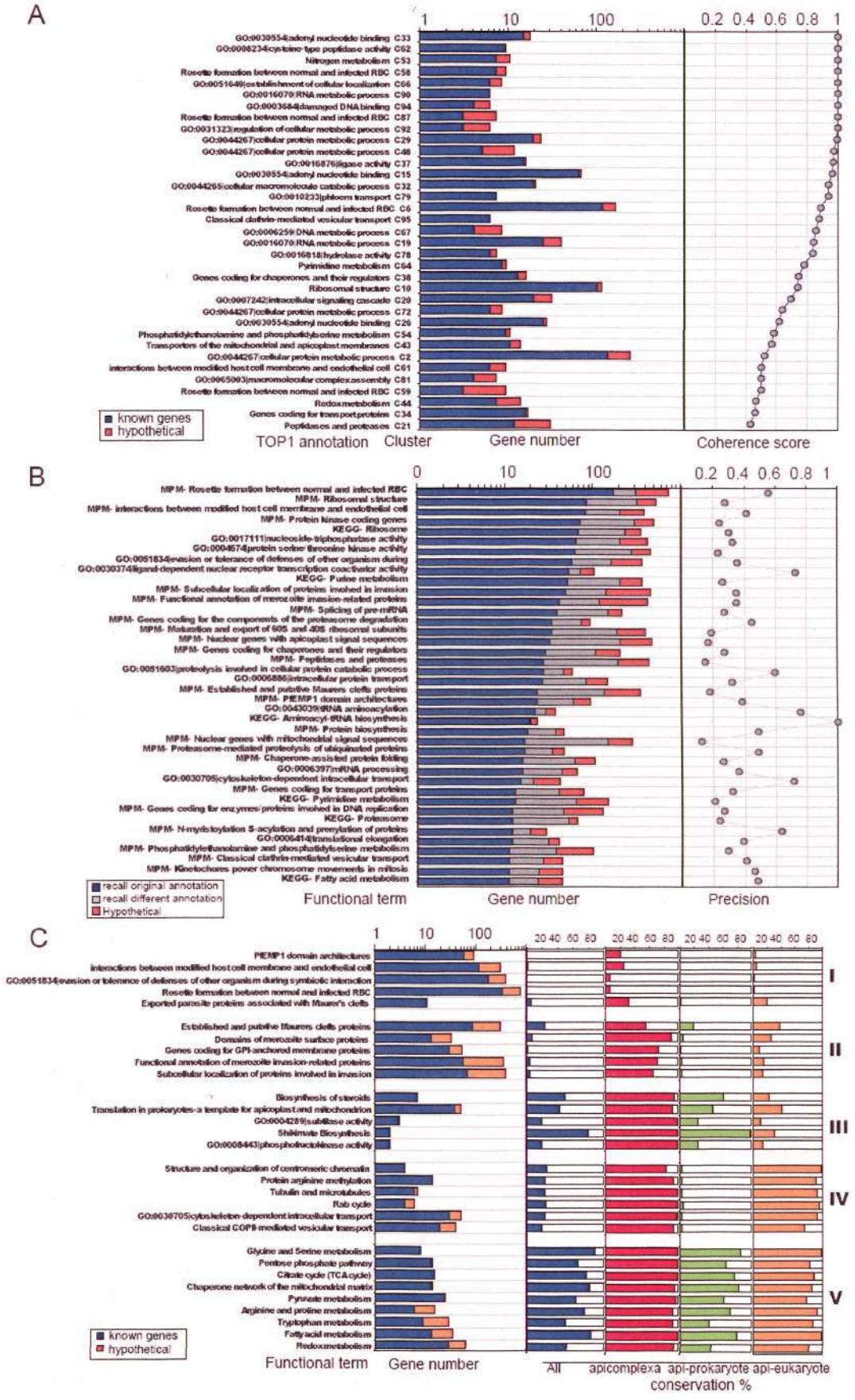


Figure 3.3. The MCL and WNC-based functional predictions and their functional categorizations. A. The bar graph summarizes the number of functionally

annotated (blue bars) and hypothetical (orange) genes in the 35 modules detected by the MCL method in the 90 % confidence network with coherence score ≥ 0.4. The modules are ordered according the coherence score and the most represented KEGG, GO or MPMP functional term is indicated. For all 105 modules with functional annotations see Figure S8. B. Bar graph summarizes the WNC based functional annotations and the leave one out analysis. For 10 functional categories with more than 10 genes with recalled original annotation ("recalled original annotation" - blue bars), the bar graph also indicate the number of genes for which the recalled annotations do not match the original ("recalled different annotation" – grey bars) and the number of hypothetical genes. The prediction precision rate calculated as the ration between the number of genes with "recalled different annotation" and genes with "recalled original annotation" is indicated in the left panel. For full list see figure S3.10. C. The conservation of different functional pathways across 210 genomes including 155 prokaryotes, 6 apicomplexa and 49 other eukaryotes is summarized and indicated for selected functional gene groups (for full list see figure S3.11). The conservation of each pathway is calculated independently as the fraction of the number of species containing potential homologs (reciprocal BLASTP hit, E-value ≤ 10^{-10}) according to four categories with: total 210 genomes (the second panel, blue bar), apicomplexa (third panel, red bar), prokaryotes plus apicomplexa (forth panel, green bar) and eukaryote plus apicomplexa (right panel, orange bar). Pathways were classified into five categories with: genes specific to *P. falciparum* (cluster I), genes conserved in apicomplexa (II), genes conserved in apicomplexa and prokaryotes (III), genes conserved in apicomplexa and other eukaryotes (IV) and genes conserved in all 210 genomes (V). The total number of functionally characterized and hypothetical genes in each category are displayed similarly in panel A.

level) and MPMP (121 terms) databases. First, the 2187 genes with functional assignments were used to evaluate the predictive accuracy of the WNC approach

using a "leave-one-out" analysis combined with top k prediction (Deng, et al., 2003). Based on this test, we could recall 2121 (97%) genes from which 996 (47%) matched the original annotations ("recalled original annotation") while for 1125, WNC derived annotations that differ from the original ("recalled different annotation") (figure 3.3B, S3.8 and S3.9). The 47% precision rate achieved by this analysis was comparable to interactome analyses of well studied model organisms including yeast, (Groth, et al., 2008; Kim, et al., 2008; Pena-Castillo, et al., 2008; Tian, et al., 2008). In comparison, an identical WNC analysis of PlasmoMAP could recall only 31% of the original annotations for the 91% of input genes (data not shown). Given the high accuracy of the WNC approach, we generated functional predictions for 2545 hypothetical proteins (95% of the total hypothetical proteins in the PlasmoINT network). The newly annotated genes could be assigned to 216 functional terms (out of the 330) with at least 5 genes with recalled original annotations (figure S3.10). For each of the 227 functional groups we also estimated the predictive precision rate as a function of the ratio between the number of genes with "recalled original annotation" and the genes with "recalled different annotation" (figure S3.10). The top 35 functional groups with the highest number of genes with "recalled original annotation" included several well defined cellular functionalities such as aminoacyl-tRNA biosynthesis (precision rate (p.r.) = 1.0), protein biosynthesis (0.49), cytoskeleton-dependent intracellular transport (0.71), clathrin-mediated vesicular transport (0.47) and fatty acid metabolism (0.47) (figure 3.3B). In addition, high precision rate for functional predictions were also

achieved for several *Plasmodium* specific functional groups linked with immune evasion and cytoadherance such as "rosette formation between normal and infected RBC" (p.r. = 0.55) and "interactions between modified host cell membrane and endothelial cell" (0.41), and invasion such as "subcellular localization of proteins involved in invasion" (0.35) and "components of the linear motor responsible for merozoite motility in invasion" (0.45) (figure 3.3B and S3.10). The high precision rates for these functional terms provide high confidence functional prediction for the hypothetical genes assigned to these functionalities and thus firm candidates for new molecular factors that are essential for growth, development as well as virulence of *P. falciparum* parasites.

Taking advantage of the phylogenetic profiles, we evaluated evolutional conservation of the functional gene groups by scoring the number of orthologues (reciprocal BLAST-based E-value > 10^{-10}) in 210 genomes of the *P. falciparum* genes in each functional category (figure S3.11). This information helped to evaluate the biological significance of different biological processes that are facilitated by these gene groups and thus their relevance for parasite growth and development. Only a small number of functional gene groups are exclusive to *P. falciparum* and show low-to-no sequence homology with known genes in other organisms including the related apicomplexan species. These include functionalities associated with *P. falciparum* virulence including host cell interaction and rosette formation (figure 3.3C cluster I). The main components of these functional groups are the subtelomeric gene families encoding several classes

of surface antigens such as *var* genes (host cell interactions), *rifin* and *stevor* (rosette formation). In addition, genes encoding proteins that are associated with Maurer's clefts and are essential for transport of parasite derived host cell surface antigens are classified by the MPMP into two functional terms: "Established and putative Maurer's clefts proteins" (figure 3.3C cluster II) and "Exported parasite proteins associated with Maurer's clefts' (figure 3.3C cluster I). While both groups exhibit minimal levels of conservation amongst prokaryotic and eukaryotic species the latter groups are highly specific to *P. falciparum* compared to the first which is moderately conserved amongst apicomplexans. Several members of both groups were recently reported to be essential for export of *P. falciparum* antigens to the surface of the infected red blood cell (Maier, et al., 2008). The newly annotated genes provide new candidates for further studies of this unique mechanism that is essential for the interaction of *Plasmodium* parasites with its host. Moreover, assessing the evolutionary conservation of the individual members might help to understand specificities of antigenic variation between different *Plasmodium* species.

Functional assignments associated with parasite invasion dominate the functional cluster that is highly conserved amongst apicomplexan but diverse from all other eukaryotic and prokaryotic species (figure 3.3C cluster II). Functionalities associated with merozoite invasion are believed to be amongst the most promising target areas for new malaria intervention strategies using both vaccine and chemotherapy approaches (Cowman and Crabb, 2006). A large number of

functionally uncharacterized genes that were assigned to this functional cluster might provide excellent targets for these efforts. To evaluate the utility of the new gene annotation we explore this gene groups further (see below). Cluster III (figure 3.3C) depicts several plasmodial functions that have a prokaryotic origin but are underrepresented in eukaryotes. These include steroid (isoprenoid) biosynthesis, apicoplast and mitochondrial translation and three homologues of subtilisine proteases (Dahl, et al., 2006; Ralph, et al., 2004; Yeoh, et al., 2007). All three functionalities are presently under consideration as potential drug targets and thus a better understanding of their biological function might contribute to this effort. In addition to these, two essential enzymes of the shikimate pathway (PFI1100w , Para-aminobenzoic acid synthetase and PFF1105c, chorismate synthase), and two enzymes associated with phosphofructo kinase activity (PFI0755c, PF11_0294) are of prokaryotic origin and are represented in the PlasmoINT network (figure 3.3C).

Nonetheless, the vast majority of the *P. falciparum* functional pathways is highly conserved in eukaryotic species or across all living organisms (figure S3.11). Figure 3.3C depicts several example pathways involved in cellular architecture, trafficking as well as maintenance of chromosomal DNA and replication that *Plasmodium* shares with the majority of eukaryotic species (Cluster IV). Pathways of basic metabolic processes that are conserved across all living organisms include the Citrate (TCA) cycle, fatty acid and amino acid synthesis, and redox metabolism (Cluster V). These data suggest that despite the extensive

diversity of the genome sequence, the majority of biological functions associated with basic metabolism as well as eukaryotic cell organization are well preserved in *P. falciparum* (figure S3.11). This also suggests that some of the molecular factors of these basic pathways represent evolutionarily diverse proteins and thus suitable targets for malaria intervention strategies (Kato, et al., 2008; O'Donnell and Blackman, 2005; Ward, et al., 2004; Yeoh, et al., 2007) . The newly annotated genes using the WNC method could present such proteins.

3.3.5 Invasome of *P. falciparum* merozoite

For further validation we chose to explore the assembled network to identify genes associated with merozoite invasion, one of the most promising targets for new malaria intervention strategies. Merozoite invasion is a complex, multiple-step process during which the parasite attaches to an erythrocyte, reorients itself and subsequently, via active penetration, enters the cell. Although more than 50 proteins were previously linked with it, the gaps remaining in our understanding of the molecular mechanisms that facilitate the invasion process indicate that many more are involved (Cowman and Crabb, 2006; Haase, et al., 2008; Soldati, et al., 2004). Using the 90% confidence interactome network, we constructed a merozoite invasion sub-network of proteins by retrieving all genes directly linked to 25 previously established invasion associated proteins (figure 3.4A and table S3.2). This sub-network contains a total of 2417 linkages connecting 418 proteins including 155 with a predicted function and 263 hypotheticals. Interestingly, the

vast majority of genes previously associated with invasion were present in this sub-network, including 43 out of 56 proteins previously predicted to be localized in intracellular compartments associated with merozoite invasion using their transcriptional and structural features (Haase, et al., 2008). The overall topological structure of the invasion sub-network overlaps well with the general view of molecular functionalities involved in invasion. It distinguishes four molecular mechanisms previously linked with it: apical organelle proteins, GPI-anchored/peripheral surface proteins, actin-myosin motors and signal transduction proteins (figure 3.4). Of the 263 hypothetical proteins that were represented in the invasion subnetwork 230 (87.5%) were also predicted by the WNC functional predictions to be involved in merozoite invasion by at least one of the top terms.

In order to evaluate the in silico predictions, 35 proteins with a high probabilistic score to be invasion related by the WNC and/or represented in the subnetwork were fused with GFP (at the C-terminus) and expressed ectopically in *P. falciparum* under the control of an appropriate promoter (Treeck, et al., 2006). The selection of this screen was biased towards proteins with predicted signal peptides (22 out of 35 proteins), given the importance of secreted proteins in the host cell invasion, immunity and their defined subcellular localisation within the apical area or the surface of the merozoite. Western Blot analyses using GFP antibodies confirmed expression of each fusion product in the transfected parasites and subsequent live cell imaging allowed their subcellular localisations. Eight

GFP-fusion proteins localized to the ER and were excluded from further validation, because accidental ER retention of certain proteins due to fusion with GFP is a known problem and therefore might not represent the true localization of the endogenous protein (Treeck, et al., 2006). Four proteins had to be omitted from the evaluation due to a very low expression of fusion protein and/or lack of a conclusive subcellular distribution. Additional 2 hypothetical proteins could not be PCR amplified.

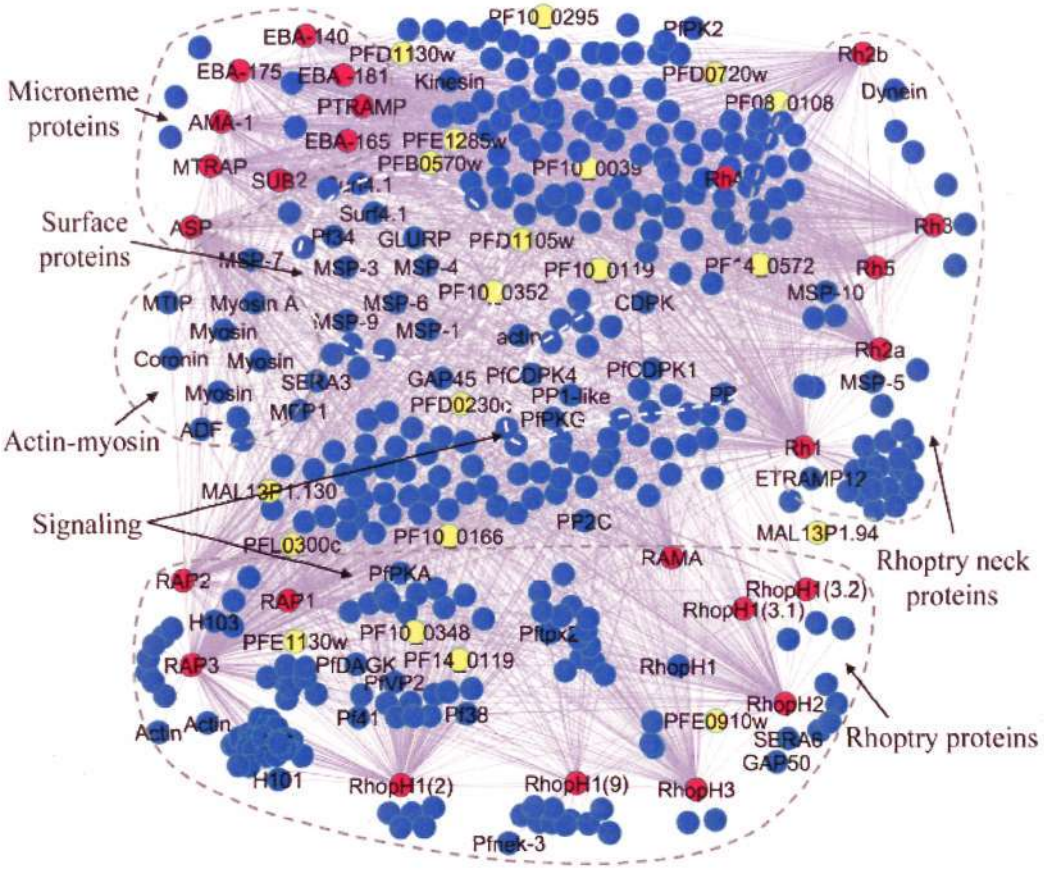


Figure 3.4 Subnetwork associated with merozoite invasion process. This sub-network has a total of 2417 links (lines) with 418 proteins (circles), and 25 core apical proteins (marked with red circles). 21 proteins (yellow circles) predicted as invasion proteins by WNC and/or present in the subnetwork were localized within the infected erythrocyte.

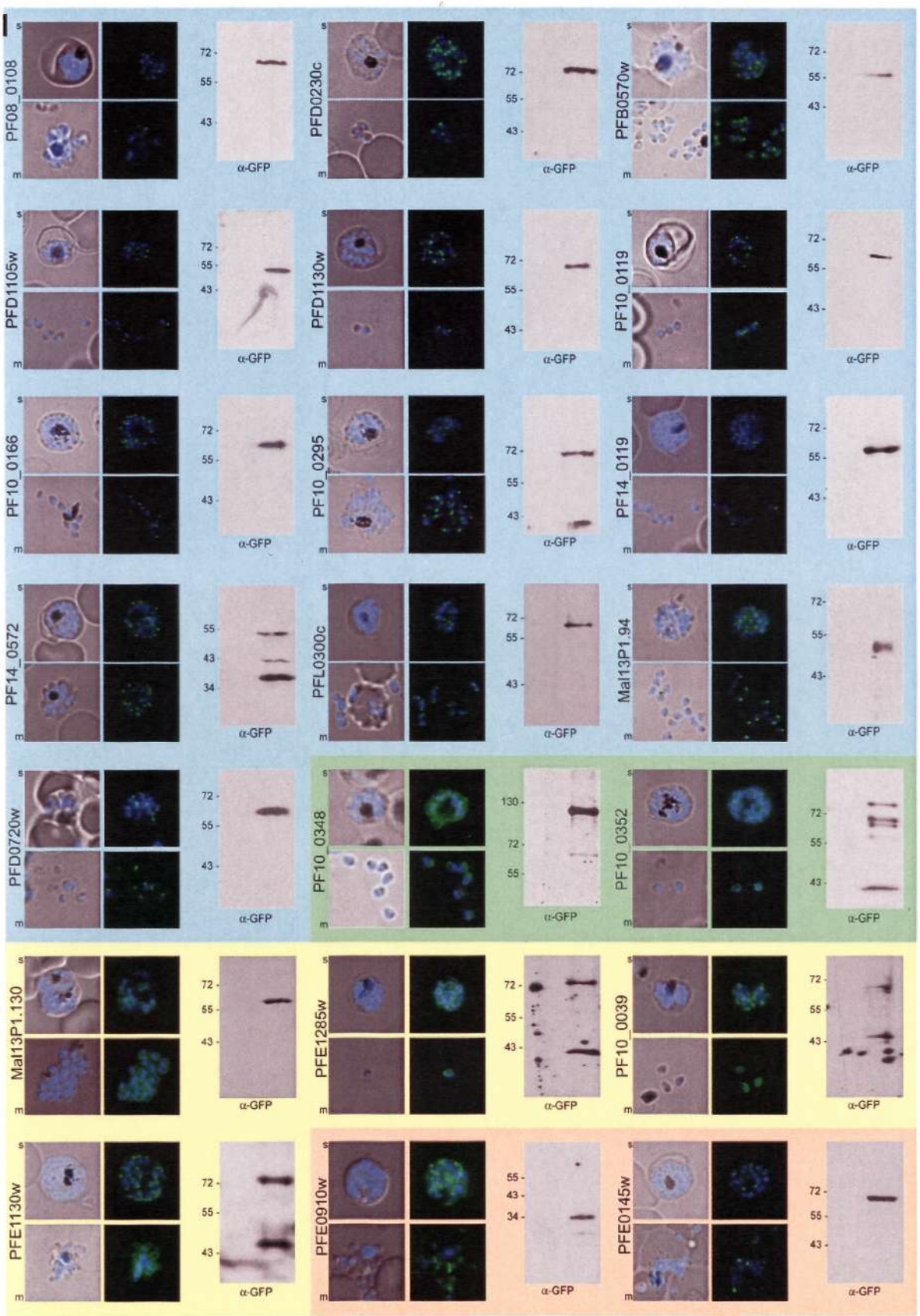
The localisation of the remaining 21 proteins led to a grouping according to the predominant localisation (figure 3.5-I). The largest group consisted of 13 proteins that showed an apical distribution of the fluorescence signal in maturing schizonts and in free merozoites after rupture (figure 3.5 II). As an example PF10_0166-GFP was co-localized with the microneme marker protein EBA-175 and showed a similar distribution (figure 3.5 II, A-B). Interestingly, the apical group also included one protein that lacked a classical N-terminal signal sequence (PFD0720w). In addition to acetylation motifs, this protein possesses an Armadillo/beta-catein like repeat that is known to be involved in protein-protein interactions. Coincidentally, in addition to its predominant apical foci, PFD0720w (and PFD1130w) also showed a faint but distinct peripheral distribution (figure S3.12 and data not shown).

The second group was represented by 2 proteins (PF10_0352 and PF10_0348) with high homology to proteins of the merozoite surface protein super-family. These proteins showed a merozoite surface distribution that was confirmed for PF10_0352 by co-localization with MSP-1 (figure 3.5I and II C.D). The third group containing four proteins (MAL13P1.130, PFE1285w, PF10_0039, and PFE1130w) exhibited a staining pattern that was reminiscent of the inner membrane complex (IMC) (Baum, et al., 2006). The IMC is tightly associated with the plasma membrane and represent a prerequisite for the structural integrity and motility of invasive parasites (Baum, et al., 2006; Baum, et al., 2008; Morrissette and Sibley, 2002). All four proteins showed a similar dynamic during merozoite

maturation as depicted for MAL13P1.130-GFP (figure 3.5 II E and S3.12): in early schizonts these proteins are present in a cramp like structure (figure 3.4 II E7) at the apical tip of forming merozoites. This structure develops into a ring like configuration (E8) before the fluorescence started to be equally distributed in the periphery of the nascent merozoite (E9-10). In this group only PFE1130w displayed a classical signal peptide.

The last group comprising PFE0910w and PFE0145w might represent the only false positive within the validated group of proteins with a localisation to either the mitochondrium or the apicoplast (figure S3.12). Interstingly, although PFE0145w has a TOP 3 prediction to be involved in invasion, it is not retrieved by this functional subnetwork.

In summary, 19 out of 21 selected proteins are associated with the structures known to be directly involved in invasion. It demonstrates that the functional predictions based on such approaches can lead to the identification of new putative targets for malaria intervention strategies.



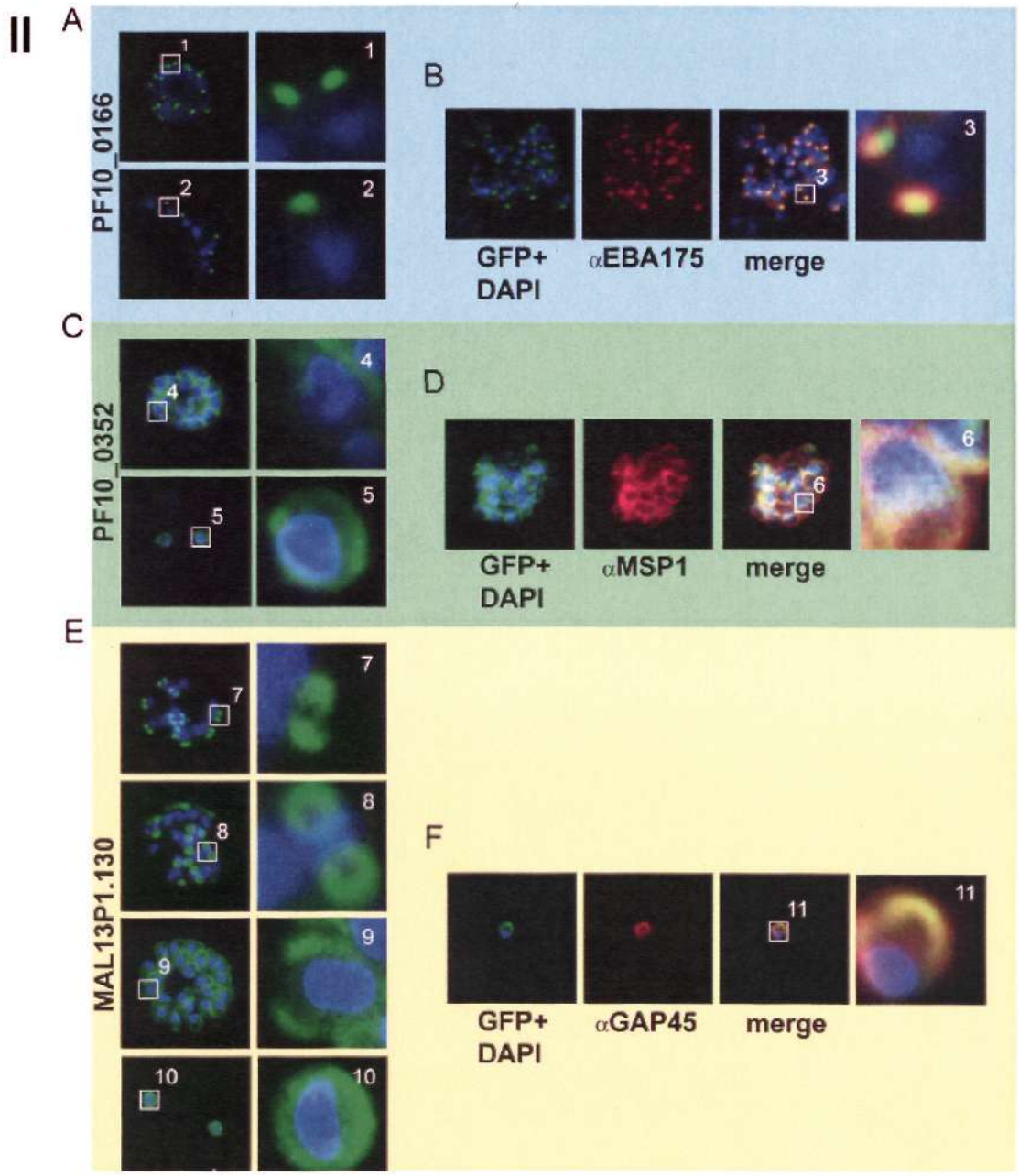


Figure 3.5 Functional analyses of merozoite invasion proteins. I. Subcellular distribution of 21 predicted proteins determined to be involved in invasion using GFP tagging; All proteins were localized in schizonts (s) and free merozoites (m). 13 proteins (PF08_0108, PFD0230c, PFB0570w, PFD1105w, PFD1130w, PF10_0119, PF10_0166, PF10_0295, PF14_0119, PF14_0572, PFL0300c, MAL13P1.94 and PFD0720w) showed a predominantly apical GFP distribution (green) and are boxed in blue. Two proteins (PF10_0348 and PF10_0352) revealed merozoite surface localisation (boxed in green) and four proteins (Mal13P1.130,

PFE1285w, PF10_0039 and PFE1130w) represented the IMC compartment (boxed in yellow). 2 proteins (boxed in orange) localized either to the apicoplast (PFE0910w) or the mitochondrion (PFE0145w). Nuclei were stained with DAPI (blue). Expression of the GFP fusion was also verified by Western-Blot analysis depicted besides each panel. II. Localisation of the apical protein PF10_0166-GFP, the surface proteins PF10_0352 and subcellular distribution and dynamics of the IMC protein MAL13P1.130; (A-B) PF10_0166-GFP (green) localized to the apical region of schizonts (s) and free merozoites (m) in unfixed (A) and fixed (B) parasites and partially co-localized (B) with the microneme protein EBA175 (red). The boxed regions are depicted in higher magnification and labelled with numbers. The nucleus is stained with DAPI (blue). (C-D) PF10_0352-GFP (green) localized to the surface of schizonts and free merozoites in unfixed (C) and fixed (D) parasites and co-localized (D) with the surface protein MSP-1 (red). (E-F) Dynamics of MAL13P1.130-GFP (green) during schizogony in unfixed parasites (E): in early schizont the MAL13P1.130-GFP emerged as a cramp like structure (enlargement E 7) at the apical tip of forming merozoites. This structure develops into a ring like configuration (E 8) before it starts to be equally distributed within the periphery of the nascent merozoite (E 9-10). (F) The MAL13P1.130-GFP co-localised with the IMC protein GAP45 (F, red) in fixed parasites.

3.4 Discussion

3.4.1 Global transactional responses of *Plasmodium* parasites to growth perturbations

Up until now, the significance of transcriptional regulation of *P. falciparum* in responses to growth perturbations remains a controversial issue. Extensive analyses of the primary sequence of proteins deduced from the *P. falciparum* genome detected only one third of transcription related factors compared to a

typical eukaryotic organism (Coulson, et al., 2004). These findings led to a suggestion that expression of the majority of *Plasmodium* proteins is regulated post-transcriptionally. Two following studies further supported these original predictions. First, exposure of *P. falciparum* cells to inhibitors of the folate synthesis pathway lead to only translational up-regulations of protein targets known to interact with these inhibitors (Nirmalan, et al., 2004). Second, treatment of *P. falciparum* cells with chloroquine, as well as one specific PKC inhibitor, resulted only in a non-specific and low amplitude transcriptional response. These finding suggested that during evolution, *Plasmodium* parasites lost some of their potential to alter their mRNA expression levels in response to variable growth conditions. Interestingly, similar results were observed for another important human pathogen, *Mycobacterium tuberculosis*, in which a number of genes essential for survival in the host lost their responsiveness to changing growth conditions and are transcribed constitutively (Rengarajan, et al., 2005). This phenomenon was attributed to the fact that this pathogen is fully adapted to its host environment and is never exposed to other types of growth conditions. Although this might be also partially true for *P. falciparum*, the strong transcriptional changes to several perturbations used in this study indicate that some pathways retained their links to transcriptional regulations and that a certain degree of flexibility exists for the parasite to respond to changing growth conditions.

Even some of the low amplitude transcriptional changes are likely to reflect physiologically relevant responses. Oakley *et al* demonstrated that a 2-3-fold

decrease in mRNA abundance of several genes of the ubiquitin-proteosome pathway resulted in approximately 15-fold decrease of overall protein ubiquitination activity in *P. falciparum* cells exposed to febrile temperatures. In our analyses, we find that even the subtle transcriptional changes are highly reproducible and dose dependent (see chloroquine treatment figure 3.1A). Moreover, the contribution of these subtle changes to the interactome networks improves the functional prediction scores based on the transcriptional co-regulation pretense (data not shown). Taken together, these data strongly indicate that despite the initial skepticism, transcriptional profiling is a suitable type of analysis for high throughput gene annotation in *P. falciparum*.

3.4.2 Gene functional network reconstruction of *P. falciparum*

In silico modeling of generic genomic systems demonstrated that even a small number of perturbations can significantly improve the confidence and gene coverage of an interactome network as long as these perturbations affect mRNA levels of 50-60% of genes in the genome (Khanin and Wit, 2007). In our analyses, we incorporated 2567 genes (48% of the genome) that exhibit at least two sequential 2-fold changes in at least one time-series. As indicated in Figure 1D, the perturbation analyses increased the likelihood score prediction values by approximately ~10-fold compared to the *P. falciparum* IDC transcriptome alone. This also led to improvements of the overall confidence and the proteome coverage of the whole interactome network compare to the identical network assembled with

the IDC transcriptome alone (Figure S3.13). In the 50% precision rate network, the average connectivity between genes was 95 using the IDC transcriptome only, but dropped to 70 when the perturbation data was incorporated (data not shown). Hence, the perturbation data are likely to be the main source of the improvements of this interactome network compared to the previously reported interactome PlamoMAP, mainly by eliminating false positive results (table 3.2 and figure S3.5 and S3.6).

3.4.3 Exploring gene function from the predictions and interactome

For 2545 hypothetical proteins of *P. falciparum,* functions were assigned based on the local network "environment" of the probabilistic interactome using the WNC method. To further validate these predictions, we manually inspected several molecular mechanisms that overlap the functional groups. Structural feature of some of the newly annotated genes provided further evidence for the precision of the WNC annotations. The first example represents 11 genes associated with the process of histone acetylation. Excluding 5 known genes: PF10_0078 (histone deacetylase), PFF0865w (H3), PF11_0062 (H2b), PFF0860c (H2a), PF11_0061 (H4), the other six hypothetical proteins (PFL1645w, PFL0635c, PFA0510w, PFF1440w, PF14_0724 and PFI1530c) contained a bromodomain motif, which is found in many chromatin associated proteins and can interact specifically with acetylated lysines (Dhalluin, et al., 1999; Hayes and Hansen, 2002; Jeanmougin, et al., 1997). 10 of the 11 proteins are present in the 90% precision rate network,

and direct linkages to these proteins identified more proteins associated with this process. These include two other histone deacetylase genes PFI1260c and PF14_0690 linking to PF10_0078, PfGCN5 (histone acetylase, PF08_0034), three additional histone genes (PFC0920w, H2a; PF07_0054, H2b; PFF0510w, H3) and a homologue of ASF1 (chromatin assembly protein, PFL1180w).

The second example is the DNA mismatch repair system where we identified 11 hypothetical proteins associated with this process together with 12 proteins previously implicated in the this process (7 DNA repair proteins: MAL7P1.206, PF00_0002, PF11_0184, PF14_0254, PFB0265c, PFE0270c and MAL7P1.145; two DNA polymerase proteins: PF10_0165 and PF10_0362; PF11_0282, deoxyuridine 5'-triphosphate nucleotidohydrolase; and PF10_0080, endonuclease). Interestingly, for 5 of the 11 hypothetical proteins, the PFAM searches performed in this study identified domains that are consistent with their involvement in the DNA repair mechanism. PF14_0051 belongs to the MutS family which is a DNA mismatch repair protein (Obmolova, et al., 2000). PFL0230w has CMP/dCMP deaminase and zinc-binding domains needed to catalyze the hydrolysis of cytidine into uridine. It has been speculated that this enzyme may be associated with the replication fork during DNA synthesis (Mathews, et al., 1988; Moore, et al., 1993). PFL1360c has one leucine-rich repeat (LRR) domain and it is involved in a variety of biological processes including DNA repair (Kobe and Deisenhofer, 1994). PF14_0538 is a protein containing a STAG domain which is typically found in subunits of the cohesin complex

(Ellermeier and Smith, 2005). Finally, PF13_0080 has a domain of RNA-directed DNA polymerase.

3.4.4 Functional analysis of new invasion proteins

The functional network predicted to power red blood cell invasion encompasses 418 proteins. Based on the invasion sub-network constructed with the 90% confidence gene linkages, we initially selected 35 predicted proteins for intracellular localization in order to validate their putative involvement in merozoite invasion. This resulted in 21 transgenic parasites line with an evaluable GFP distribution within the infected erythrocyte comprising 14 proteins with a classical signal peptide, 2 proteins with a putative signal anchor and 5 proteins without any apparent localisation motif. 13 proteins revealed an apical GFP localisation reflecting the initial biased selection for the functional screen towards protein with predicted signal peptides. The apical area is defined by its associated secretory organelles (rhoptries, micronemes and dense granula) as depicted in figure 3.5 (II, B3) using EBA-175 as a microneme marker protein. These organelles compile an unknown number of secreted proteins that play not only an important role for host cell interaction, but are highly interesting for vaccine and drug development (Cowman and Crabb, 2006). For instance, PFD0230c (also known as DPAP3) contains a serine protease domain and was recently identified in a forward chemical genetic screen as one of the key regulators for merozoite egress (Arastu-Kapur, et al., 2008). This function is in a good agreement with the

network based assignment as well as the localization studies that demonstrated PFD0230c to be transported to the apical organelle(s) (figure 3.5 II). PF08_0108, an aspartate protease (also known as plasmepsin X), belongs to the *P. falciparum* specific plasmepsin family with 10 members highly homologous to pepsinogen A (Coombs, et al., 2001). While some of the plasmepsins are involved in hemoglobin degradation in the digestive vacuole, PF08_0108 is expressed in late schizonts and localizes as an ectopically expressed GFP fusion protein to the apical organelles of merozoites (figure 3.5 I). This suggests that plasmepsin X has evolved a distinct role in the Plasmodium life cycle, being involved either in merozoite egress, or invasion. Two additional proteins from the apical group, PFB0570w and PFD1105w were previously described as rhoptry proteins with adhesive properties. PFB0570w (PfSPATR, secreted protein with altered thrombospondin repeat) displays a degenerated TSP-domain with a multi-stage expression profile (Chattopadhyay, et al., 2003) and PFD1105w (PfAARP, asparagin rich parasite protein) was shown to bind to surface structure on the erythrocyte (Wickramarachchi, et al., 2008). Again, these functions are in a good agreement with the network based assignment.

The two proteins that were indentified to been surface located are both encoded by genes within the *msp* cluster on chromosome 10, that encodes multiple proteins belonging to the merozoite surface super family. Noteworthy, PF10_0348, which is located in an msp cluster on chromosome 10, encodes an additional DBL domain, which might mediate initial receptor binding with the host erythrocyte

(Wickramarachchi et al manuscript submitted). MAL13.P1.130, PFE1285w, PF10_0039 and PFE1130w showed a subcellular localization at the IMC similar to GAP45, a member of the actin-myosin motor, a complex that plays a crucial role in invasion (Baum, et al., 2006). MAL13P1.130 was initially characterized as a 6 transmembrane protein by a proteomic approach characterizing detergent resistant membrane fractions of parasites in schizont stages (Sanders, et al., 2005), while PFE1130w represents a 7 transmembrane domain IMC protein. Structurally distinct without any transmembrane domain (or signal peptide) are PFE1285w and PF10_0039 that belong to the family of alveolins and are known to play a structural role in IMC architecture (Baum, et al., 2008).

It will be crucial to further validate new proteins from the predicted invadome to deepen our understanding of the invasion process on the molecular level, although only functional studies will provide a basis for rational drug and vaccine development.

3.5 Concluding remarks and outlook

Human malaria remains one of the most dangerous infectious diseases in the world affecting 300-500 million and killing 1-2 million people each year. The fast spreading resistance to the majority of the available chemotherapeutic agents and the lack of an operational vaccine create a serious health concern for the future. Better understanding of the *Plasmodium* parasite biology and especially functional relevance of the numerous *Plasmodium* hypothetical genes is paramount for the

development of new malaria intervention strategies. Here we carried out extensive transcriptional profiling of *P. falciparum* responses to chemically induced growth perturbations and used these data to reconstruct a gene interactome network that allow us to predict function of 2550 *Plasmodium* hypothetical genes. The high accuracy of these predictions was demonstrated by the functional validation of 21 selected protein candidates from which 19 localized to intracellular compartments associated with the invasion machinery. Given the low efficiency of all available reverse as well as forward genetics in *P. falciparum* (Balu, et al., 2005; O'Donnell, et al., 2002), these data demonstrate that transcriptional profiling of growth perturbations provides a powerful technique for functional genomics of *Plasmodium* parasites. The functional predictions based on this network provide highest precision rates compared to the presently available interactome network. All these gene predictions are available in an online database on the following website: http://zblab.sbs.ntu.edu.sg/network/index.html.

Although this network covers 88% genome of *P. falciparum* and about 2550 hypothetical proteins were assigned function clues, to improve the accuracy of the network and predictions of gene function, more data will be incorporated in the future. For example, more growth perturbation data, expression profiles from mutant strains and proteomic expression profiles are being investigated. Currently, 20 proteins were chosen for intracellular localization analysis to evaluate the network-based predictions of gene function. These invasion proteins would be further analyzed to illustrate the molecular function in the invasion

process in the future, which would promote deep understanding of the mechanism of merozoite invasion in *P. falciparum*. Furthermore, more proteins would be selected for function analysis, such as proteins associated histone modification and DNA binding proteins.

3.6 Materials and methods

3.6.1 *Plasmodium falciparum* genome

The 4.4 version genome, 5363 protein encoding open reading frames and 53 pseudo genes of *P. falciparum* are downloaded from PlasmoDB (http://www.plasmodb.org/download/) excluding mitochondrion and plastid genes. The current annotation of CDSs is based on the 5.4 version genome. All linkages and calculations of genome coverages are based on this gene set.

3.6.2 Parasite culture, treatment and microarray and plasmid transfection

Cells of *P. falciparum* strain 3D7 and Dd2 were grown and maintained as previously described (Trager and Jensen, 1997). Growth assay of each drug or compound were performed in 2% hemotocryt with 5% parasitemia at one particular stage. Parasitemia of new rings of next cell cycle were counted to calculate the inhibitory concentration at 50% (IC_{50}). Parasites were treated with appropriate drug or compound concentrations ($IC_{50/90/180}$) and collected in a course with 5-8 time points taken in regular time intervals (30-120minutes). Genome-wide gene expression profiling was conducted using a long

oligonucleotide representing all 5363 *P. falciparum* genes, and the microarray hybridizations were carried out as previously described (Hu, et al., 2007). *P. falciparum* asexual stages (3D7) were transfected as described previously (Fidock and Wellems, 1997). Positive selection for transfectants was achieved using 10 nM WR99210.

3.6.3 Gene expression profiles

Currently, mRNA expression profiling represents one of the most extensive functional genomics data sets, and it has been proved that genes with similar expression profiles are more likely to be co-regulated, functionally related and encode interacting proteins (Bhardwaj and Lu, 2005; Eisen, et al., 1998; Ge, et al., 2001; Tornow and Mewes, 2003). Till now, only several data sets about *P. falciparum* are available, and most of them are data type of life cycle. Here we collected data of 247 micrarrays from different drug or inhibitor treatment (table 3.1), together with 42 microarray experiments of cell cycle from lab or field strains and from the published data. Data of each drug/inhibitor experiment were extracted from NOMAD database, and each gene profile was represented the average intensities of all oligos that map to that gene. The missing data were fixed by K-nearest neighbor method in R package (Troyanskaya, et al., 2001). Gene profiles were assembled when existing in all experiments.

Pearson Correlation Coefficient (PCC) between the expression profiles across the entire perturbation experiments panel for each gene pair were calculated

to evaluate the complexity of the growth perturbations. Based on the benchmark data (described in the following section), we calculate a likelihood score as a function of a ratio of the probabilities of positive and negative observations for different PCC thresholds to systematically evaluate functional relationships of transcriptional co-regulated genes in the growth perturbation data.

To construct the gene functional network we also incorporated others datasets, the IDC transcriptome datasets of 3D7, Dd2 and HB3 (148 microarray experiments) (Llinas, et al., 2006). To indicate the strength of functional association of each gene pair by gene expression profiles, PPCs were calculated independently across each dataset first and a called "optional average" method was used to average the three PPCs as the final correlations (fPPC). Briefly, Fisher's z-transform (David, 1949; Huttenhower, et al., 2006) was used to average two PPCs from two independent IDC transcriptomes and compared to the PPC from perturbation data. If the later is smaller, the final PPC is the PPC from perturbation data. Otherwise, the final PPC is equal to the average PPC from three datasets using the Fisher's Z transform. To illustrate the advantage of this method, we reconstructed the network using the average PPCs by Fisher's z-transform of all gene pairs. We divided the final correlations into 19 bins. For each bin we assessed its overlap with the benchmarks (table S3.1).

3.6.4 Protein-protein interaction data based on yeast two-hybrid experiments

Physical protein-protein interactions (PPI) reflect functional associatioin of the

corresponding genes in most, if not all, cases. As for direct experimental observations of protein interacrions in *P. falciparum*, a set of 2811 interactions among 1308 proteins that generated by the application of yeast two-hybrid method was used (LaCount, et al., 2005). We defined all interactions as only one bin and assessed its overlap with the benchmarks (table S3.1).

3.6.5 Protein-protein interaction data based on domain-domain interaction evidences

The basic units of proteins are domains and proteins interaction with each other through their domains. Bioinformatics methods are developed to predict the domain interactions by integrating the experimental data sources of protein-protein interactions from different species as well as other data sources (Deng, et al., 2002; Lee, et al., 2006; Riley, et al., 2005; Sprinzak and Margalit, 2001). Lee et al. (Lee, et al., 2006) predicted a set of high-confidence domain-domain interactions by integrating multiple biological data sets from four species (yeast, worm, fruit fly and human). We mapped this data set to *P. falciparum* to predict the protein-protein interactions based on these domain interaction evidences. Briefly, first we predicted the domain information of all malaria proteins using HMM method based on the PFAM database (Sonnhammer, et al., 1998), and generated all possible protein domain-domain pairs and offered the confidence score (likelihood score) of each domain pair. The score of a domain pair was assigned to a pair of proteins containing the domains. If different scores existed between a pair proteins

arising from different interacting domain pairs, the maximum of the scores was assigned to the pair. We divided the confidence scores into 6 bins. For each bin we assessed its overlap with the benchmarks (table S3.1).

3.6.6 Prediction of functional linkages using phylogenetic profiles

Phylogenetic profiles of all proteins were calculated using the mutual information method (Date and Marcotte, 2003). Briefly, the protein sequences of *Plasmodium falciparunm* were compared with reference organisms (210 reference organisms, including 155 prokaryotes and 55 eukaryotes, were downloaded from the NCBI and the ENSEMBL) using BLASTP (Altschul, et al., 1997). For each protein i, a vector was generated with elements p_{ij}, where $p_{ij} = -1/\log E_{ij}$. Here we predetermined the E-value threshold is equal to 1e-4 according to the prediction power of different E-value thresholds (Sun, et al., 2005). That is, $p_{ij} = 1$ when the E-value is greater than or equal to the predetermined E-value threshold. As a metric of phylogenetic profile similarity, the mutual information was calculated between pairs of phylogenetic profiles (Date and Marcotte, 2003; Sun, et al., 2005). In practice, mutual information is calculated on histograms of p_{ij} values, binned in 0.01 intervals, with resulting MI values ranging from 0−2.4. To avoid the effects of paralogs in *Plasmodium*, we deleted all protein pairs originated from paralogous proteim pairs, in which paralogs were defined by their BLASTP E-values (<= 1e-15). Figure S3.2 showed the hierarchical clustering of phylogenetic profiles of all *P. falciparum* proteins and an example of correlated proteins. We divided

mutual information scores into 15 bins. For each bin we assessed its overlap with the benchmarks (table S3.1).

3.6.7 Reference and benchmark sets

For the validation and prediction of protein-protein functional relationship, we need to have reference datasets to serve as gold-standards of positives and negatives. The Kyoto-based KEGG database (Kanehisa, et al., 2002) provides metabolic and regulatory pathway annotation for genes. Previous studies have proved the KEGG database to be an excellent reference set for evaluating functional linkages (Date and Marcotte, 2003; Date and Stoeckert, 2006; Lee, et al., 2004; Marcotte, et al., 1999). KEGG maps about 10% genes of *P. falciparum* into at least one pathway or cellular systems (examples including "glycolysis", "ribosome", "proteosome"). We extracted 71 pathways in which it has at least two genes, including total 492 genes. The KEGG database produces original 12,493 positive gene pairs, finally 11,046 positive pairs was determined after kicking out any pairs in which gene participate more than 3 pathways, thus avoiding promiscuous members. A total of 61,721 negative gene pairs were created according to all possible genes pairs based on all genes in the 79 KEGG pathways (8 pathways only have one gene) excluding all original positive pairs and any pairs in which they share any GO terms up to 4^{th} level in all three GO categories (Ashburner, et al., 2000). Table S3.1 showed the parameters of nave Bayesian network of all datasets based on this reference dataset.

In order to test the predictive values of the input data, we assemble a positive benchmark dataset that comprises of 11,046 linkages between 492 *P. falciparum* genes that fall into 71 distinct KEGG pathways. We also assemble a negative benchmark dataset that contains 61,721 gene pairs that do not fall into a common KEGG pathway and do not share a Gene Ontology (GO) term up to 4th level. For the phylogenetic profiling, domain-domain interaction and transcriptome datasets, there are positive trends between the linkage evidence values (such as PCC) and the benchmark based likelihood scores (figure S3.3). This data suggest that the calculated likelihood scores reflect the functional relationships between *P. falciparum* genes and are applicable as input values for assembly a probabilistic interactome network.

3.6.8 Integration of the data sets by Bayesian probabilistic model

Bayesian model are efficient to integrate heterogeneous data for the task if combining evidences (Date and Stoeckert, 2006; Jansen, et al., 2003; Lee, et al., 2004; Troyanskaya, et al., 2001). Four dataset types are evaluated by the standard positive and the negative benchmarks, and each gene pair was assigned one likelihood score. Then all likelihood scores given by different data types are integrated by the Bayesian model, which generates the final prediction score for a potential protein linkage based on a product of the likelihood scores from each of the four data sets with no penalties for missing evidence from any set.

$$\text{Likelihood Score (LS)} = LS_{PPC} \times LS_{PHY} \times LS_{PPI} \times LS_{Domain}$$

PPC is gene expression profile linkages. *PHY* is phylogenetic profile linkages. *PPI* is experimental protein-protein interaction linkages. *Domain* is domain-domain interaction linkages. The posterior probability was computed based on Bayesian formula, $O_{posterior} = O_{prior} \times \text{LS}$.

3.6.9 Cross validation of the integration results

We performed a 10-fold cross-validation to evaluate the overall performance of the prediction. Briefly, first the positive and negative benchmarks were randomly divided into ten separate equal sets, and nine of them were used as the training set to calculate the likelihood scores and the remaining one set as the test to identify the positive s and negatives. We ran this process ten times so that each of the ten sets was a test set and the remaining nine constituted the training set. Finally, all true positives (TP) and false positives (FP) were summed up under different likelihood score cutoffs to evaluate the ratio of true positives to false positives. The positive predictive value (PPV, TP/(TP+FP)) was also calculated as the fraction of true positives to the total number of true positive and false positive.

3.6.10 Characterization of the network structure and identification of local modules among the network

We used several essential variables (node degree, degree distribution, clustering coefficient) to characterize the overall topological structure of the network (Barabasi and Oltvai, 2004). Node degree (connectivity), k, is the number of links

that the node has to other nodes. The degree distribution, P(k), gives the probability that a selected node has exactly k links, which allows us to distinguish between different classes of network. When the degree distribution approximates a power-law, $P(k) \sim k^{-\lambda}$, it means the network is scale-free. Clustering coefficient characterizes the tendency of nodes to form clusters or groups. If the distribution of clustering coefficient, C(k), follows $C(k) \sim k^{-1}$, the network structure is hierarchical (Ravasz, et al., 2002).

We searched the local modules in the network using Markov Cluster (MCL) algorithm which is a fast and scalable unsupervised graph clustering algorithm (Enright, et al., 2002; Krogan, et al., 2006). Comparing analysis of Markov Clustering (MCL) and other methods concludes that MCL performs robustly and superiorly to extract protein complexes from interaction networks (Brohee and van Helden, 2006). To define the parameter of granularity, we followed the method of Wuchty and Ipsaro (Wuchty and Ipsaro, 2007) by optimizing the functional coherence and size of the clusters (Lee, et al., 2004). The networks and sun-networks were laid out and visualized using Cytoscape 2.5 (Shannon, et al., 2003).

3.6.11 Network-based gene function prediction

3.6.11.1 Weighted neighbor counting method

We used one neighbor counting method weighted by the likelihood score because the likelihood score of each linkage could represent the functional similarity

between two proteins.

$$f(i, j) = \sum LS(m)\, \delta(j) / \sum LS(m)$$

where the f(i,j) is the propability of gene i having function j. The LS(m) is the likelihood score of the m^{th} neighbor of gene i. $\delta(j)=1$ if the gene has function *j*, else $\delta(j)=0$. Without threshold, we assigned an unannotated protein with k functions having the top k statistic scores (see the following). The performance of the predictions is evaluated by plotting precision against recall over vary thresholds as adopted in Deng et al (2003). For a given threshold β, precision and recall are defined as:

$$\text{Precision} = \sum^{V}_{i} k_{i,\beta} / \sum m_{i,\beta} \qquad \text{Recall} = \sum^{V}_{i} k_{i,\beta} / \sum n_i$$

where n_i is the number of known functions of protein i; $m_{i,\beta}$ is the number of functions predicted for protein i at threshold β and $k_{i,\beta}$ is the number of functions predicted correctly for protein i at threshold β. V is the set of all functionally known genes.

3.6.11.2 TOP k statistics

In this analysis we utilize the top k predictions statistics in which top k number are evaluated simultaneously for the final functional prediction. To determine the optimal k, we compare the prediction precision and the sensitivity of all annotation terms for every gene when k is equal 1, 3 and 5 regardless the thresholds of the prediction scores. Top 3 assignments regardless the prediction scores had overall 50% predictive precision at 85% recall of the total annotated genes in the network and had significant improvement than the overall 42% predictive precision of top 1

assignment. Although top 5 assignments had overall 52.6% predictive precision, the sensitivity was only 23% at the same recall comparing to 29% of top 3 assignments (figure S3.8A). When the threshold of prediction score was set at 0.14, top 1 assignment had 50% predictive precision recalling for 64% of the total annotated genes in the network (figure S3.8B). Based on this threshold, top 3 assignments had overall 54.6% predictive precision and top 5 assignments had overall 55.4% predictive precision recalling 64% of annotated genes in the network (figure S3.8). Comparatively, at the same recall, the predictive precision rate of top 3 assignments was 58% regardless the threshold of prediction score (figure S3.8). Taken together, we define the k-value equal to 3 to ensure the largest genome coverage and highest predictive precision rate in the overall gene function prediction.

3.6.11.3 Comparison with other methods

I. chi-square approach (Hishigaki, et al., 2001).

$$S(i, j) = \frac{(n(i, j)-e(i, j))^2}{e(i, j)}$$

The n (i, j) is the number of proteins interacting with protein i and has function j. The e(i, j) = #Nei(i) * π_j is the expected number of proteins in its all neighbors having function j, where #Nei(i) is the number of neighbor proteins of protein i.

II. FS weighted average method (Chua, et al., 2006)

FS weighted average method is one neighbor counting method considering both direct and indirect neighbors based on the functional similarity distance, which was calculated according to the method adopted by Chua et al. 2006.

$$FS(i, j) = \frac{\sum^{u}(S(i, u)\,\delta(u, j) + \sum^{v} S(u, v)\,\delta(v, j))}{\sum^{u}(S(i, u) + \sum^{v} S(u, v))}$$

FS (i, j) is the predicting score having j function for gene I by FS weight average method. The S(i, u) is the FS-weight score of u^{th} neighbor of gene i. The FS-weight score (S) was calculated based on the formula 5 of Chu et al. (2006), in which the evidence score of each linkage was defined by LS/(LS + 1). $\delta(j)=1$ if the gene has function j, else $\delta(j)=0$.

Several computational methods mentioned above were used to predict gene function and test the predictive accuracy. Compare to neighbor counting method, Chi square approach and FS-weight average method, the weighted neighbor counting method had a significantly higher overall prediction precision rate regardless of the thresholds (figure S3.9).

3.6.12 Genes associated with the invasion subnetwork and experimental validation

The 25 apical proteins (table S3.2) locate at the apical organelles (microneme, rhoptry and rhoptry neck) were taken as the core hubs (Cowman and Crabb, 2006) to build the merozoite invasion sub-network by retrieving all direct linkages to the hubs. 20 uncharacterized proteins in the sub-network were selected to confirm the results.

3.6.13 Nucleic Acids, Antisera and Immunoblots

Genes of interest were either amplified using gDNA or cDNA derived from 3D7

parasites. PCR was carried out using cDNA gene specific primers summarized in Tab. SX. PCR products were digest with *Kpn*I and *Avr*II and ligated into the transfection vectors pARL$_{ama\text{-}1}$-GFP (Struck, et al., 2005) or pARL$_{ama\text{-}1}$-TY1 that encode a C-terminal GFP or TY1 tag. To ensure late expression the AMA-1 promotor is used to drive transcription (Treeck, et al., 2006). Proteins from late stage parasites were separated on 10% SDS-PAGE minigels and immunoblots were performed and developed as previously described (Struck, et al., 2005). Anti-GFP (Roche) or anti-TY1 (Diagenode) was used as a primary antibody and sheep anti mouse IgG horseradish peroxidase (Roche) was used as a secondary antibody.

3.6.14 *Immunofluorescence and analysis of GFP expressing parasites*

Images of unfixed GFP-expressing parasites were observed and captured using a Zeiss Axioskop 2plus microscope, a Hamamatsu Digital camera (Model C4742-95) and OpenLab software version 4.0.4 (Improvision Inc.). DNA was stained with DAPI (1:1000, Roche).

Chapter 4 Global gene expression of *Plasmodium falciparum* in response to protein kinase or phosphatase inhibitors

4.1 Summary

Protein kinases and phosphatases play important roles in the development of malaria parasites. Inhibitors of both functionalities have been shown to block invasion by *P. falciparum* but their mechanisms of actin remain largely unknown. Here we study the effect of several classes of inhibitors of protein phosporylation pathways including CaM (W7), calcium/CaM-dependent protein kinase (KN93), myosin light chain kinase (ML7), conventional multiple protein kinases (staurosporine) and two calcineurin inhibitors (FK506 and CsA) on erythrocyte invasion by *P. falciparum*. The main goal is to further define the prospective modes of action of these inhibitors by analyzing the gene expression response. First, the growth assays show that these inhibitors effectively inhibit erythrocyte invasion by *P. falciparum in vitro* with the parasite cells arrested in the late schizont stage. Second, the global gene expression profiling using a genome-wide *P. falciparum* DNA microarray shows that these inhibitors induce diverse but specific transcriptional responses when parasite cells were treated at the early schizont stage. Interestingly, several transcription factors and signaling genes were up-regulated by the inhibition of calcium dependent signaling and calcineurin signaling pathways, which suggests that the phosphorylation

and/or dephosphorylation play vital roles in the gene expression regulation in *P. falciparum*.

4.2 Introduction

Malaria parasites have a complex life cycle that includes a sexual development in the mosquito vector, exo-erythrocytic cycle and intra-erythrocytic developmental cycle (IDC), in the liver and the blood of the human host respectively. From these however, the IDC is responsible for all clinical symptoms and it is also a target for the vast majority of the malaria intervention strategies. Several large scale transcriptome analyses of the *P. falciparum* life cycle uncovered a broad transcriptional regulation that controls expression of the vast majority of the genome. It was shown that each cellular pathway is timed to a specific stage of the *Plasmodium* life cycle in a "just in time manufacturing" fashion (Bozdech, et al., 2003; Le Roch, et al., 2003). But little is known about the exact timing of regulation of gene expression in *P. falciparum*. Modulation of protein phosphorylation through the antagonistic effects of protein kinases and protein phosphatases is a major regulatory mechanism of most cellular processes in eukaryotic cells. Sequence analysis of the genome identified tens of protein kinases in *P. falciparum* including many homologues conserved in other eukaryotic signaling proteins (Gardner, et al., 2002; Ward, et al., 2004). Over the past few years several genes encoding *Plasmodium* protein kinases have been characterized and show that some protein kinases are expressed in specific stages (Bozdech, et

al., 2003; Le Roch, et al., 2003) and play important roles in the development of the parasite (Billker, et al., 2004; Canduri, et al., 2007; Doerig, et al., 2002; Kappes, et al., 1999). Previous studies has also shown that *Plasmodium* protein kinases diverge significantly on both structure and function from their homologs in other eukaryotes (Doerig and Meijer, 2007; Ward, et al., 2004). Recently, it was shown that a set of genes involved in protein modification particularly protein phosphorylation are regulated by a transcription factor with an AP2 domain (De Silva, et al., 2008). Taken together, these studies lend a hope that specific inhibition of parasitic kinases is achievable and may lead to the development of novel control agent against malaria. However, validation of a given kinase as a drug target requires strong evidence that its activity is essential for parasite growth and/or differentiation. It also requires understanding how these kinases integrate in the cellular machinery of the *Plasmodium* cells and what roles they play in gene expression during the parasite development.

Several protein kinase or phosphatase inhibitors have been shown to have a similar blocking effect of erythrocyte invasion by *P. falciparum* compared to several protease and cytokinesis inhibitors (Dluzewski and Garcia, 1996). In particular, W7, calmodulin (CaM) antagonist, was shown to block the erythrocyte invasion presumably via Ca^{2+} depletion and a subsequent affect on a putative calcium/calmodulin-dependent signal pathways (Vaid, et al., 2008; Ward, et al., 2004). KN93, a specific inhibitor of CaM kinases, interacts with the calcium/CaM-binding domain of CaM kinases (CaMK) to inactivate these kinases

(Means, 2000). In *Plasmodium*, KN93 was shown to block the formation of ookinetes from zygotes (Silva-Neto, et al., 2002) and as well as the gamete formation by blocking calcium-dependent protein kinase 4 (PfCDPK4) (Billker, et al., 2004). Another calcium/calmodulin dependent kinase, myosin light chain protein kinase (MLCK), can be specifically inhibited by ML7. Currently no studies have been completed to characterize its inhibitory effect on *P. falciprum*. Saturosporine, an inhibitor of serine/theronine kinases effectively inhibits the erythrocyte invasion of *Plasmodium* (Dluzewski and Garcia, 1996; Ward, et al., 2004). Xestoquinone is a Pfnek-1 inhibitor with in vitro antimalaria activity but little is known about its mechanism (Laurent, et al., 2006). A competitive inhibitor of cAMP (cAMP Rp-isomer), which inhibits cAMP-dependent protein kinase (PKA), also results in inhibition of apical regulated exocytosis in sporozoites and hepatocyte infection (Ono, et al., 2008). HDTAB (hexadecyltrimethylammonium bromide) inhibits Choline kinase (PfCK) in a dose-dependent manner and offers very potent antimalarial activity against *P. falciparum* (Choubey, et al., 2007). The phosphatase inhibitor of okadaic acid has a strong inhibitory effect both on invasion and development of *P. falciparum* (Dluzewski and Garcia, 1996). Two calcineurin inhibitors, cyclosporine A (CsA) and FK506, have been proven to inhibitor the erythrocyte invasion by *P. falciparum* (Bell, et al., 1994; Kotaka, et al., 2008; Kumar, et al., 2005) and calcineurin had the contrary effect of kinase and was able to dephosphorylate proteins in *P. falciparum* (Dobson, et al., 1999; Kumar, et al., 2005). Although all the above-mentioned kinase or phosphatase inhibitors

have effective anti-malarial effects, the mechanisms of blocking the erythrocyte invasion are largely unknown.

The aim of this study is analyze the genome-wide transcriptional response of *P. falciparum* to several classes of protein kinase inhibitors in order to further understand a role of their corresponding the signaling pathways in the progression of the IDC. For these studies, we chose a selection of inhibitors targeting CaM (W7), calcium/CaM-dependent protein kinase (KN93), myosin light chain kinase (ML7), and the calcineurin pathway (FK506 and CsA) (summarized in table 4.1). In addition we include a conventional inhibitor of multiple protein kinases (staurosporine). Growth assays show these inhibitors effectively inhibit the erythrocyte invasion in vitro by *P. falciparum*. Global gene expression profiling on a genomic scale using microarray technology shows these inhibitors have diverse transcriptional responses when parasite cells were treated from the early schizont stage and specific gene responses induced by inhibiting classic signaling pathways are observed. Interestingly, several transcription factors and signaling genes were up-regulated induced by the inhibition of calcium dependent signaling and calcineurin signaling pathways, suggesting the phosphorylation and/or dephosphorylation play vital roles of the gene expression regulation in *Plasmodium falciaprum*.

Table 4.1 Inhibitors associated with protein kinases

Inhibitors	IC50 (nM)	Possible primary gene target	Possibly involved in functional pathway
ML-7	1224	MLCK	Myosin-actin
W-7	1272	CaM	Calcium/CaM-dependent pathway
KN-93	1232	CaMKII	Calcium-dependent signaling pathway
Staurosporine	80	various PKs	Multiple signaling pathways
Cyclosporine A	88	Cyclophilin A	Calcineurin pathway
FK506	118	FKBP56	Calcineurin pathway

4.2 Results and discussion

4.2.1 Effects of kinase and calcineurin inhibitors on the development of *P. falciparum*

We studied the effects of protein kinase inhibitors on the progression of the malaria parasite intraerythrocytic developmental cycle (IDC). Using growth inhibitions assay we show that all utilized inhibitors (ML7, W7, KN93, staurosporine, cyclosporine A and FK506) inhibit the growth of *P. falciparum* when the parasites were treated at the early schizont stage. The 50% inhibitory concentration (IC50) was determined individually for each compound for the identical culturing conditions (2% hemotocryt with 5% parasitemia, see materials and methods). Three calcium dependent signaling inhibitors (ML7, W7 and KN93) have similar IC50s (aproximately 1.2uM) while the IC50 concentrations of calcineurin inhibitors FK506 and CsA are 118nM and 88nM, respectively (table 4.1). The conventional multiple target kinase inhibitor, staurosporine, has the

lowest IC50 concentration (80nM), which suggests the its high potency to inhibit *Plasmodium* growth.

To better understand the events of development of *P. falciparum*, synchronous cultures (Dd2 strains) were treated with double 90% inhibitory concentrations concentration (IC90) at early-stage schizonts over 32 hours. The effect of the inhibitors on the parasite morphology was monitored by Giemsa-stained smears prepared with the treated *P. falciparum* cells collected at selected time-point intervals (4hr, 8hr, 14hr and 32hr) (figure 4.1). In the ML7, W7 and KN93 treated cultures, nearly all parasites developed into morphologically normal merozoite clusters associated with hemozoin particles at time 8hr post treatments (similar to the untreated control culture). Arrested mid-schizonts were observed in cultures treated with staurosporine and two calcineurin inhibitors (CsA and FK506) as early as 8 hours after the inhibitor treatment. Interestingly, at time 14hr post treatment, clusters of merozoites loosely distributed in the infected erythrocytes were observed in both calcium-dependent inhibitors (ML7, W7 and KN93) and calcineurin inhibitors (cyclosporine A and FK506) treated cells. This is a sharp contrast to the untreated control culture in which all parasites re-invaded new erythrocytes and entered the ring stage (figure 4.1). 32hr treatments of *P. falciparum* cultures with these inhibitors resulted in the cell death which is characterized by dense small cell bodies with contracted nuclei (figure 4.1). These data show three calcium-dependent inhibitors (ML7, W7 and KN93) blocked the rupture process of mature schizonts and two calcineurin inhibitors

(cyclosporine A and FK506) had similar effects on cell development but appeared to interfere with the cell development already in the mid-schizont stage. Comparatively, conventional kinase inhibitor of staurosporine had much more serious effects on the parasite development (figure 4.1). It is consistent with the broad specificity of staurosporine inhibiting multiple kinases through the prevention of ATP binding to the kinase. As shown in figure 4.1, this causes severe toxicity to immature schizonts.

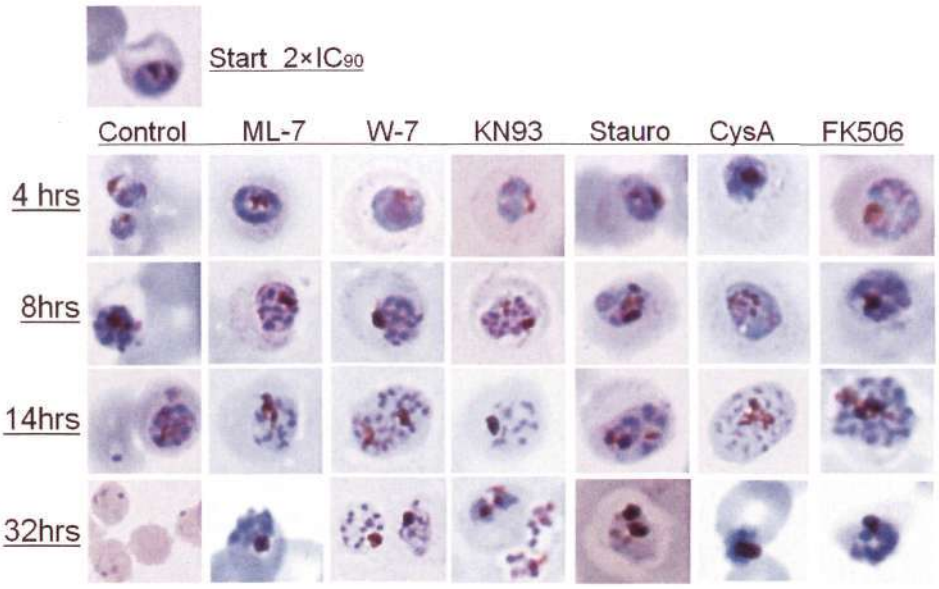


Figure 4.1 Effects of protein kinase and calcineurin inhibitors on *P. falciparum* development. Parasite morphology was monitored by Giemsa stain microscopy 4, 8, 14, and 32 h after addition of inhibitor. During the first 8 h (schizont stage development), no significant morphological differences were observed between the treated and untreated parasites. During their subsequent development, the untreated controls progress to the next generation (formation of ring stages) while the treated cell remains arrested at the late schizont stage. The appearance of dense black shrunken cells 32 h post-treatment is consistent with parasite death.

In *Toxoplasma gondii*, Ca2+ release from intracellular stores governs tachyzoite egress, micorneme secretion, motility and host invasion (Carruthers and Sibley, 1999; Kieschnick, et al., 2001; Lovett and Sibley, 2003; Moudy, et al., 2001) and increasing edvidence suggests that similar mechanisms operate in invasion stages of *Plasmodium* (Gantt, et al., 2000; Kawamoto, et al., 1993). Taken together, these data suggest that calcium dependent signaling (calcium, calmodulin and calcium/calmodulin-dependent protein kinase) is essential to activate the rupture of the schizonts and suppress of calcium signal interferes with the rupturing process.

4.2.2 Gene expression response to kinase and calcineurin inhibitors in *P. falciparum*

To analyze the gene expression response to the protein kinase and calcineurin inhibitors in *P. falciparum*, we carried out a course with 7 time points taken in regular time intervals (1 – 2 hours) over 12 hours (1, 2, 4, 6, 8, 10 and 12hr, starting from schizont stage, around 32hpi) for each compound treatment with IC50 concentration and measured the global gene expression level with a long nucleotide DNA microarray representing all 5363 *P. falciparum* genes (Hu, et al., 2007). We extracted genes whose mRNA abundance was altered by 3-fold changes to determine the significantly expressed genes. To guarantee the continuous observation of gene expression changes through the time course, we retained the genes that at east one neighbor of the peak changed data points (>=3 fold) should

have 2 fold changes (double three-two filtering). This filtering would avoid the noise from a single experiment. Diverse transcriptional responses were observed from the three classes of inhibitors: calcium-related inhibitors (W-7, ML-7 and KN93), calcineurin inhibitors (cyclosporine A and FK506) and conventional kinase inhibitor (staurosporine) (figure 4.2A). The myosin light chain kinase inhibitor (ML7) induced transcriptional changes of 561 genes. The CaM antagonist induced noticeable expression alteration of 636 genes, and Ca/CaM-dependent protein kinase inhibitor of KN93 induced transcriptional changes of 340 genes. Two calcineurin inhibitors CsA and FK506 induced closer transcriptional changes of 291 and 285 genes, respectively. Conventional kinase inhibitor staurosporine caused a more complex response which included 570 genes (figure 4.2A). Profoundly, a huge increase of 637 genes was changed if we used double two-two filtering comparing to other inhibitors (figure 4.2B). This dramatic response is consistent with the non-selective targets of the staurosporine and the severe toxic effects on the development of parasites.

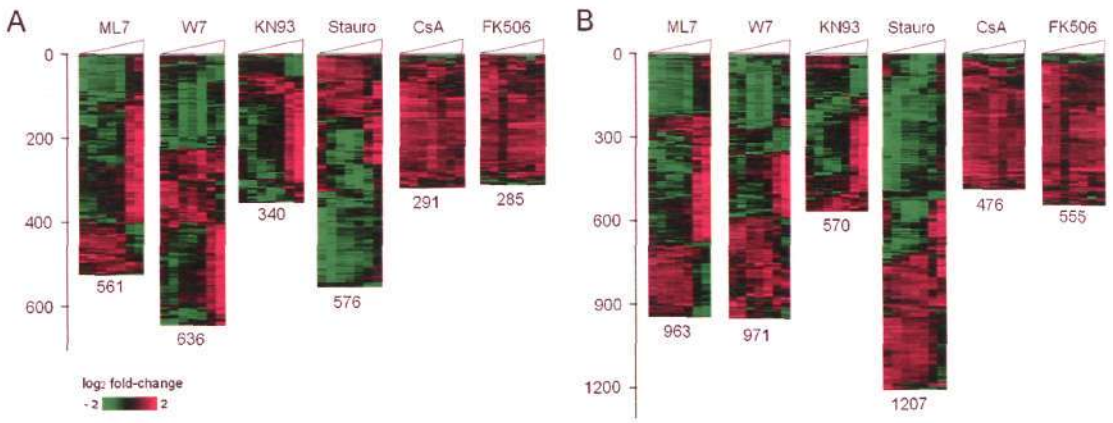


Figure 4.2 The transcriptional changes induced by compounds of ML7, W7, KN93, stausporine, cyclosporine A and FK506. In each diagram of transcriptional changes,

each line is an expression profiles for one genes. The expression profiles were transformed by the expression profiles in the untreated control cells and thus red color reflects up-regulation and green color down-regulation of gene transcription induced by the inhibitor. Gene profiles were clustered using Cluster (Eisen, et al., 1998) A. The transcriptional changes calculated from double three-two filtering method. B. The transcriptional changes calculated from double two-two filtering method.

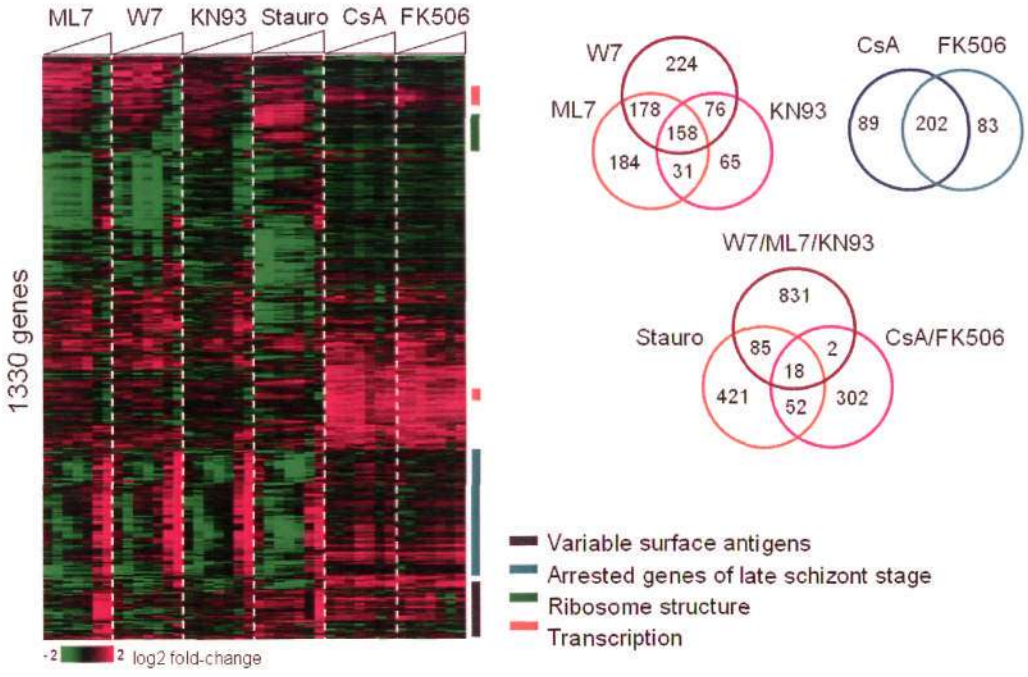


Figure 4.3 Comparative analysis of the transcriptional changes induced by compounds of ML7, W7, KN93, stausporine, cyclosporine A and FK506 based on double three-two filtering. Total 1330 gene were retained after the filtering through the six treatments (left). In the three classes of inhibitors (W7-ML7-KN93, CsA-FK506 and staurosporine), gene expression changes induced by the same class inhibitors had large overlapping and less by between different class (right).

Interestingly, there are significant overlaps between transcriptional responses induced by inhibitors that belong to the same class. CaM antagonist W7 and myosin light chain kinase (MLCK) inhibitor ML7 had similar transcriptional responses with 336 shared genes (53% of W7 and 60% of ML7) (figure 4.3). KN93 targeting calcium/CaM-dependent protein kinases also showed similar transcriptional changes to W7 and ML7. 69% of genes whose expression was affected by KN93 were olso found to be effected by ML7 and W7 (figure 4.3). All these genes that show 2-fold changes (double two-two filtering) induced by KN93 share also comprise approximately 70% of ML7 and W7 transcriptional responses. The significantly overlaps of the transcriptional responses induced by the three inhibitors is likely due to their proteins target that is shared between these compounds are likely related to calcium dependent signaling pathway. These gene responses are considerably different from those induced by two inhibitors of CsA and FK506 (figure 4.3). CsA and FK506, which inhibit the calcineurin (calcium/CaM-dependent protein phosphatase) dependent signaling pathway in eukaryotic cells (Liu, et al., 1991) by binding to cyclophilin and PfFKBP35 respectively, had significantly similar transcriptional changes (approximately 70% of responsive genes). The transcriptional changes induced by the conventional kinase inhibitor staurosporine also shared with approximately one third of the changes induced by ML7, W7 and KN93, individually. Although staurosporine had 103 and 70 responsive genes shared with ML7-W7-KN93 group and CsA-FK506 group (figure 4.3), the overlapping genes between CsA-FK506 group and

ML7-W7-KN93 group was only 20. These data show the conventional and non-selective protein kinase inhibiting activity of staurosporine and the diverse roles of protein kinase and phosphatase in gene regulation of *P. falciparum*. Interestingly, common transcriptional changes were observed through all inhibitors as well, such as translation machinery, variant surface antigens and the genes expressed in late schizont stage who were arrested by the inhibitors (figure 4.3). We also found a lot of transcription related genes (transcription and cell cycle factors) and signal transduction related genes (kinases, regulators, phophatases) were significantly changed induced by the kinase and calcineurin inhibitors.

Taken together, this data indicates the specific roles of the protein phosporylation and dephosporylation pathways on transcriptional regulation in *P. falciparum*. To understand the representations of the cellular systems and the functional roles of genes in response to the inhibitors, in the following section, we systematically analyze and discuss the gene expression changes in the context of different functional classes and different types of protein phosphorylation pathways.

4.2.3 Cell responses of variable surface antigens and ribosome structure genes

The effects of anti-malaria drugs or compounds on malaria pathogenesis are profound and not well understood. *P. falciparum* variable surface antigens (VSA) that include tree major antigen coding gene families (*var, rifin, stevor)* are considered to be major contributors to the variable nature of malaria pathogenesis.

On the microarray, 85% of VSA genes have unique probes ensuring the specific detection of their expression (Hu, et al., 2007). Previously a generalized up-regulation of VSA expression has been observed in transcriptional responses to the heat shock environment (Oakley, et al., 2007). We observe that 95 VSA genes (total 142 detected) including 58 *var* (14 pseudo or truncated), 31 *rifin* (4 pseudo or truncated) and 6 *stevor* (3 pseudo or truncated) were up-regulated in the parasites treated by kinase inhibitors and only 2 pseudo *var* genes down regulated (figure 4.4A). Especially, in ML7 treatment, most *rifin* and *var* genes were upregulated in last two time points. VSA genes may be related to cellular stress response or inhibition of kinase activities would affect the transcriptional regulation of these variable genes. Nuclear myosin plays important roles in transcriptional regulation (de Lanerolle, et al., 2005). ML7 would affect the nucleus myosin through inhibiting the myosin light chain kinase and lead to upregulation of VSA genes. This will be tested in the future.

The proteins of ribosome structure apeared to have the similar responses to anti-malaria drugs or compounds. 51 out of total 112 detected genes in genome were observed to be induced by all six protein kinase inhibitors (figure 4.4A). Interestingly, like VSA genes, most genes of ribosome structure were up-regulated (44 out of 51). These genes were not possibly arrested because the parasite development at the early treatment was not affected (figure 4.1). It suggests the control of gene expression at the translational level is an important mechanism involved in cellular stress response.

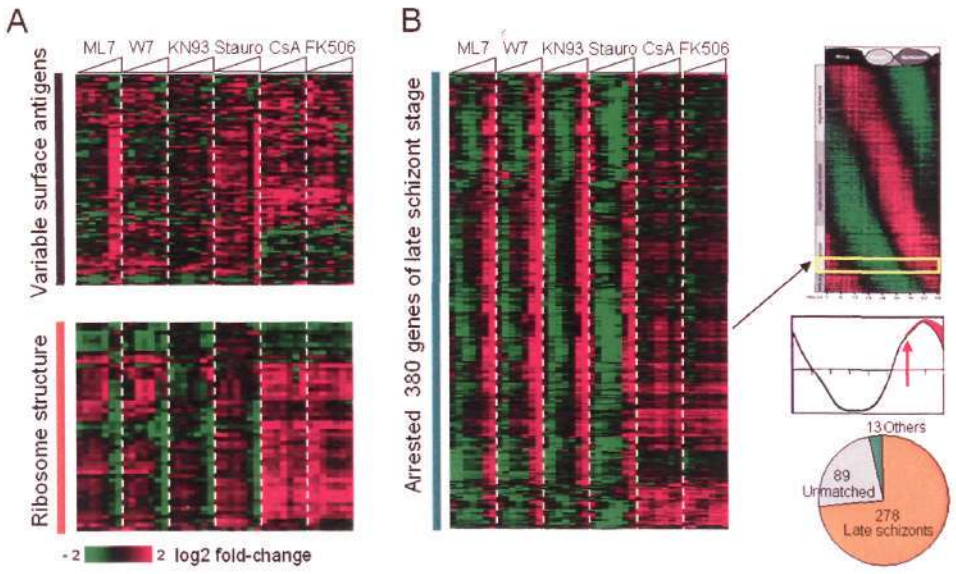


Figure 4.4 Functional analyses of the transcriptional changes induced by compounds of ML7, W7, KN93, stausporine, cyclosporine A and FK506 based on double three-two filtering method. A. Clustering diagrams of variable surface antigen genes group (up) and ribosome structure gene group (down); B. Clustering diagram of genes group of late schizont stage.

4.2.4 Gene expression in late schizont stage arrested.

Among the genes induced by four protein kinase inhibitors ML7, W7, KN93 and staurosporine, there was a large cluster of genes which are under normal growth condition specific to the late schizont stage (figure 4.3 and 4.4B). grouping particular, 278 of 380 genes induce by the compounds (74%) are also highly expressed during the late schizont stage according to the IDC transcriptome (Bozdech, et al., 2003). Most of them are suggested to be involved in the merozoite invasion process including most of the known invasion-related proteins (Cowman and Crabb, 2006; Haase, et al., 2008; Soldati, et al., 2004). The additional 89 genes

have no information in the IDC transcriptome and only 13 genes were out of this stage (6 late trophozoite stage and 7 early ring stage) (figure 4.4B). It indicated that these genes were still expressing at 10hr of post treatment, and comparatively in the untreated control parasites they gradually slow down (figure 4.4B). Interestingly, CsA and FK506 were able to induce the mRNA abundance of only 52 genes expressed during the late schizont stage as early as 1 hour after the compound was added to the culture (figure 4.4B). This includes several cytoskeleton related genes (etramp11.1 and 12, RESA, RAP3, Rh3, Clag3.1, PfCRT2, PfNBP-1) and protease (hydrolase and SUB2). These data suggest gene expression switches (regulation) in the late schizonts were interfered by ML7, W7, KN93 and staurosporine but not by CsA and FK506, although inhibition of both protein phosporylation and dephosporylation (calcineurin) pathways arrested the development of parasite cells morphologically.

3.2.5 Transcriptional changes induced by inhibitors related to calcium dependent signaling

Three inhibitors in this study are related to calcium dependent signaling (figure 4.5A). W7 is the calmodulin antagonist to adjust the intracellular calcium concentration. Calmodulin can bind to a wide array of protein kinases with calcium/calmodulin binding domain, such as CaMK and MLCK. KN93 and ML7 are specific inhibitors to CaMK and MLCK, respectively (Billker, et al., 2004; Dluzewski and Garcia, 1996; Means, 2000; Silva-Neto, et al., 2002; Vaid, et al.,

2008; Ward, et al., 2004). The early transcriptional responses of *P. falciparum* to three inhibitors (W7, ML7 and KN93) included 477 up or down regulated genes excluding all variable surface antigen, ribosome genes and genes expressed induce as a result of the developmental arrest in late schizont stage (figure 4.5B). It suggests that calcium or calmodulin dependent signaling play vital roles in the development of the *P. falciparum* life cycle and are directly linked with transcriptional regulation. This group includes 224 (47%) up and 181 (38%) down regulated genes with 160 functionally known and 317 hypothetical genes. Functional analysis shows that these genes include stress response genes (such as glutathione S-transferase), trafficking system proteins (such as Rab family genes), cell surface and adhesion proteins (such as glycophorin binding protein related antigen), signaling proteins (such as PfCDPK3), DNA replication and repair (such as replication factors), and transcription regulation (such as CCAAT-binding transcription factor). Specifically, we observed up-regulation of 5 cell cycle and transcription regulators induced by this group of inhibitors (table 4.2). Interestingly, the only one down-regulated gene (prohibitin) is a transcription suppressor that was previously found to bind proteins that belong to the family of E2F transcription factors (O'Connor, et al., 2001; Wang, et al., 1999). These proteins may be possibly involved in the process of *P. falciparum* cell cycle development and gene expression regulation. We also observed up-regulation of 5 protein kinases while only one gene (PF08_0019) encoding PfRACK (Receptors for activated C kinases) was down regulated (table 4.2). PfRACK is conserved in other

eukaryote and conspicuously spread throughout the schizont, suggesting it might play a key role in the regulatory processes of malaria parasite life cycle (Madeira, et al., 2003). Interestingly, one putative protein kinase (MAL13P1.84) was down regulated in ML7 treatment but up regulated in both W7 and KN93 treatments. This finding might indicate a difference in the effect of calcium/camodulin-CaMK and calcium/calmodulin-MLCK on gene regulation. Two phosphatase genes (PF14_0523 and PFL1260w) were down-regulated, and one phosphatase (MAL8P1.109) and one phosphatase activator (PF14_0280) were up regulated. Taken together, these data suggest that calcium and calmodulin signals can regulate gene expression through calcium/calmodulin dependent protein kinases (CaMKs and MLCK) and also that CaMKs and MLCK share downstream regulators of gene expression.

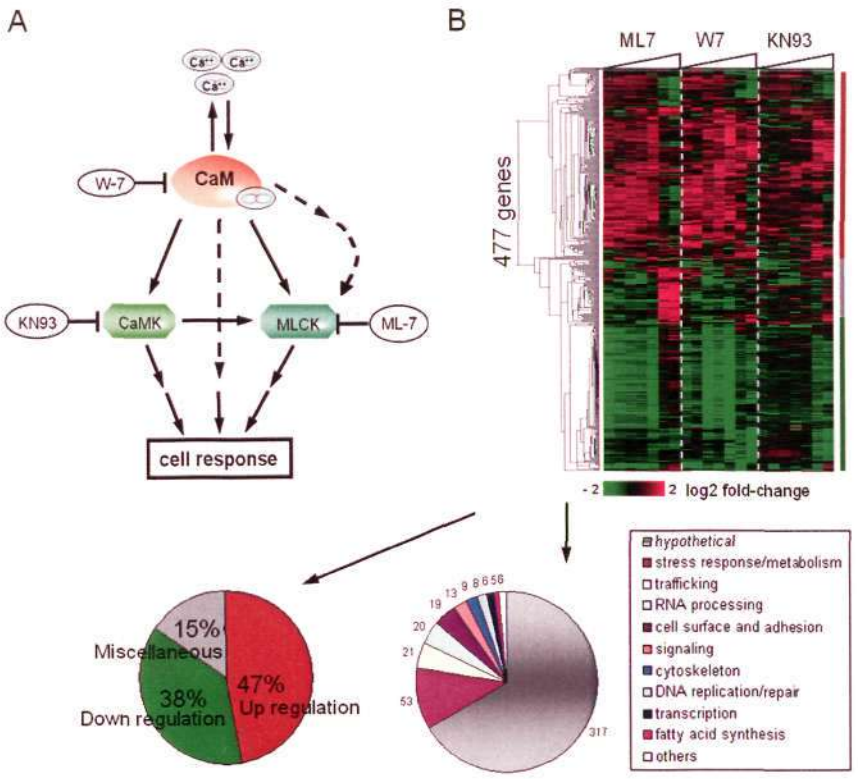


Figure 4.5 Functional analyses of the transcriptional changes induced by calcium

dependent signaling inhibitors ML, W7 and KN93 based on double three-two filtering method. A. The signaling pathways of protein targets of ML7, W7 and KN93. B. Comparative and functional analyses of the transcriptional changes induced by compounds of ML7, W7 and KN93.

Table 4.2 Transcriptional changes of gene transcription regulators and signaling factors induced by inhibitors of calcium dependent signaling.

Function	Gene ID	Gene Name	Fold changes *		
			ML7	W7	KN93
Transcription	PF13_0043	CCAAT-binding transcription factor	2.47	3.03	2.1
	PFE1470w	cell cycle regulator protein, putative	2.25	3.16	1.39
	PFL1180w	Chromatin assembly protein (ASF1)	2.86	3.74	1.6
	PFE0920c	cyclin2 related protein	3.74	4.03	2.44
	PF08_0074	DNA/RNA-binding protein Alba	2.81	3.25	1.53
	PF10_0144	prohibitin	-1.9	-5	-3.17
Signaling	PFC0420w	calcium-dependent protein kinase 3	2.23	2.44	3.5
	MAL13P1.84	protein kinase	-3.2	3.04	1.49
	PF08_0019	PfRACK	-3.57	-2.38	-2.44
	PFI0505c	selenide water dikinase	2.69	3.54	1.15
	MAL7P1.100	serine/threonine protein kinase, Pfnek-4	3.29	3.51	2.89
	PFC0060c	Serine/threonine protein kinase	2.13	3.46	3.47
	PFA0380w	serine/threonine protein kinase	3.63	4.15	3.03
	PFD1180w	trophozoite antigen r45-like protein	4.33	1.83	2
	PF11_0224	circumsporozoite-related antigen	-4.6	-5.7	-2.75
	PFL1260w	hydrolase / phosphatase	-2.1	-2.12	-3.62
	PF14_0523	protein phosphatase 2C	-3.8	-3.3	-7.37
	MAL8P1.109	Protein phosphatase 2C	3.05	4.7	4.13
	PF14_0280	phosphotyrosyl phosphatase activator	2.99	3.74	1.95

* maximum fold change in the time course

4.2.6 FK506 has a similar action mode to cyclosporine A

Compounds CsA and FK506, inhibit the calcineurin-dependent signaling pathway

in eukaryotic cells (Liu, et al., 1991). In *P. falciparum*, CsA binds to cyclophilin and FK506 binds to PfFKBP35 both to interfere with the calcineurin signal pathway by inhibiting the activity of calcineurin, a calcium/CaM-dependent protein phosphatase (Bell, et al., 1994; Kumar, et al., 2005). Global transcriptional responses of *P. falciparum* to CsA and FK506 showed CsA and FK506 induced transcriptional changes of 269 and 255 and suppressed 21 and 30 genes (>3-fold changes, double three-two filtering), respectively (figure 4.2). Interestingly, both inhibitors affected the transcription of 202 genes (69% of CsA and 71% of FK506) with 192 transcripts exhibiting a >3-fold increase in abundance (figure 4.6A). Good correlation in the global transcriptional responses induced by both inhibitors points to a similar mode of action, which is consistent with the presumed mode of action in other eukaryotic cells: suppression of calcineurin-dependent signaling pathways (Bell, et al., 1994; Kumar, et al., 2005; Liu, et al., 1991). Biological functional analysis showed 12 genes associated with cells surface and adhesion, 11 stress response or metabolic genes, 7 transcription related genes, 5 trafficking system genes and several kinase and cytoskeleton genes had 3 fold changes, although the majority of these genes are functionally uncharacterized (103 hypothetical proteins), 24 VSA genes (11 Var., 9 rifin and 4 stevor) and 29 genes of ribosomal subunits (figure 4.6B). Interesting most transcriptional responses induced by both FK506 and CsA affect the transcriptional profile of the early schizont development which underlines the eventual developmental arrest and subsequent parasite death (figure 4.2 and 4.6A). The specificity of the

FK506/CsA-induced transcriptional response is further supported by the fact that other unrelated inhibitors of cell signaling pathways, W7, ML7, KN93 and staurosporine, affected expression of distinct groups of genes with minimal overlaps (figure 4.3). The transcriptional changes of a large group of transcription related genes suggested the importance of gene expression regulation by the calcineurin pathway in *P. falciparum* (table 4.3). These include early up-regulation of several transcription factors, such as TATA-binding protein (TBP, PFE0305w), ruvB-like DNA helicase (possible TBP interacting protein, PF08_0100), Myb1 (PF13_0088) and transcription repressor high mobility group box protein (HBP, MAL8P1.72, high mobility group box domain). In the transcriptional changes induced by CsA and FK506, mitogen-activated protein kinase 2 (Pfmap-2, PF11_0147) was also up-regulated early. Interestingly, the transcription repressor HBP (MAL8P1.72), which is a possible target of MAPK (mitogen-activated protein kinase) in human (Xiu, et al., 2003), was up-regulated of more than 20 fold even in the early time points (1-2 hr) in both treatments. It suggests that calcineurin pathway can regulate gene expression through mitogen-activated kinase (MAPK) pathway. Interestingly, two up-regulated cell cycle related genes (PFE1215c, developmentally regulated GTP-binding protein 1 and PF10_0370, enhancer of rudimentary homolog) suggest the inhibition of calcineurin pathway would affect the development of *P. falciparum*, consistent with the arrested shcizont (figure 4.1). Taken together our data suggest that FK506 binding protein PfFKBP35 appears to be essential for the progression of the *P. falciparum* life cycle through the gene

expression regulations. We speculate that the mode of FK506 and its antimalarial effect may be mediated through targeting PfFKBP35 and subsequent inhibition of the parasite calcineurin.

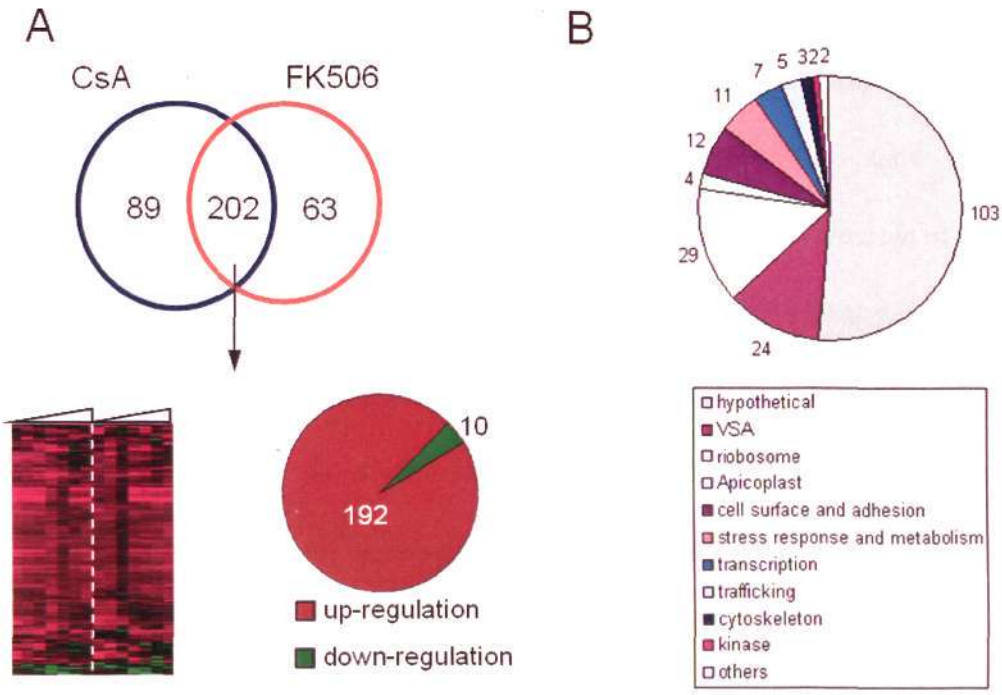


Figure 4.6. Comparative (A) and functional (B) analyses of the transcriptional changes induced by compounds of cyclosporine A and FK506.

Table 4.3 Transcriptional changes of gene transcription regulators and signaling factors induced by inhibitors of calcineurin dependent signaling.

Function	Gene ID	Gene Name	Fold changes *	
			CsA	FK506
Transcription	PFE1215c	Drg1	3.88	2.64
	PF10_0370	Enhancer of rudimentary homolog	4.28	4.38
	MAL8P1.72	high mobility group protein	29.1	22.5
	MAL7P1.151	modification methylase-like protein	3.74	3.48
	PF13_0088	Myb1 protein	4.64	4.38
	PF08_0100	ruvB-like DNA helicase	3.67	5.05
	PFE0305w	TAT-binding protein, TBP	3.67	3.89
signaling	PF11_0147	mitogen-activated protein kinase 2	3.78	3.12
	PF11_0220	protein kinase	3.37	3.85

4.3 Conclusion and outlook

Protein kinases and phosphatases play important roles in the development of the malaria parasites. We studied the effects of several classes of inhibitors of protein phosphorylation and dephosphorylation pathways on the growth of *Plasmodium* cell and the regulation of gene expression. Growth assays show that these inhibitors effectively inhibit the erythrocyte invasion by *P. falciparum in vitro* with the parasite cells arrested in the late schizont stage. The global gene expression profiling using a genome-wide *P. falciparum* DNA microarray (mentioned in chapter 2) shows that diverse but specific transcriptional responses induced by these inhibitors were observed when parasite cells were treated at the early schizont stage. Interestingly, several transcription factors and signaling genes were significantly regulated resulting from the inhibition of calcium dependent signaling and calcineurin signaling pathways, which suggests that the phosphorylation and/or dephosphorylation play vital roles in the gene expression regulation in *P. falciparum*.

Currently, we are performing real-time PCR to confirm the regulated transcription factors and signaling genes. Also we are trying to identify protein intermediates between the protein kinases and the transcription factors using a novel proteomic technique of 2D-DIGE/MS. We hope to establish the putative regulation network between protein kinases and transcription factors and their target genes to illustrate the principle of gene regulation in *P. falciparum*.

4.4 Materials and methods

4.4.1 Parasite culture

Cells of *P. falciparum* strain Dd2 were grown and maintained in a 2% suspension of purified human RBCs and RPMI 1640 media supplemented with 0.25% Albumax II (GIBCO, Life Technologies, San Diego, California, United States), 2 g/L sodium bicarbonate, 0.1 mM hypoxanthine, 25 mM HEPES (pH 7.4), and 50 ug/L gentamycin, at 37°C, 5% O_2, and 6% CO_2. Cells were synchronized by two consecutive sorbitol treatments for three generations (Trager and Jensen, 1997).

4.4.2 Growth assays and inhibitor treatments

To perform the *P. falciparum* growth assay, 1 ml of synchronized early schizont-stage parasites in hypoxanthine-free complete medium (5% parasitemia and 2% hematocrit) were added to each well in 96-well plate, and compounds were added at the first well to a final drug concentration of 8 uM and a final volume of 2 ml (add new cultures). The cultures were mixed well and took 1 ml mix from the first well to the second well and mix to dilute the compounds, and then followed to the third well and repeated 9 times. Finally compounds in 10 wells of the plate ranged from 8 uM to 15.625 nM. The plates were then incubated in a chamber with a standard gas environment at 37°C for 20 hours. After the 20-h incubation, cells in each well were made smear and stained by Giemsa. Rings were counted under

microscope. The growth assay of each compound was done in three replicates. The counts of rings were plotted against the logarithm of the drug concentration, and the curve was fitted by nonlinear regression using the formula sigmoid dose-response to calculate the inhibitory concentration at 50% (IC_{50}) (table 4.1). For the inhibitor treatments, cell cultures were carried out in the same way as the growth assays and cells were treated by the compound at the concentration of its IC50. After 1, 2, 4, 6, 8, 10 and 12 hour treatment, cells were collected and washed with pre-warmed PBS, and flash-frozen in liquid nitrogen and stored in -80°C for RNA isolation.

4.4.3 RNA preparation and cDNA labeling

P. falciparum RNA sample isolation, cDNA synthesis, labeling, and DNA microarray hybridizations were performed as described by Bozdech et al. (Bozdech, et al., 2003). Samples for individual timepoints (coupled to Cy5) were hybridized against a reference pool (coupled to Cy3). The reference pool was comprised of RNA samples from 3D7 strain representing all developmental stages of the parasite. For this pool, sufficient cDNA synthesis reactions were performed for all hybridizations, and then all reference pool cDNAs were combined into one large pool and then split into individual aliquots for subsequent labeling and hybridization.

4.4.4 Microarray manufacture, hybridization and scanning

Microarray manufacturing and hybridizations were conducted as previously described (Bozdech, et al., 2003; Bozdech, et al., 2003). Briefly, all oligonucleotides in 384 well plates were printed on the polylysine-coded glass slides using BioRad microarray printer system. Printed slides were post-processed by rehydration, UV cross-linking and succinic anhydride (ALDRICH, Cat. 239690) block. The labeled cDNA samples were hybridized to the chip in MAUI system (BioMicro, Utah, United States) for 12-14 hours at 65°C. Data were acquired and analyzed by GenePix (Axon Instruments, Union City, California, United States). Array data were stored and normalized in Acurity 4.0 system (Axon Instruments, Union City, California, United States).

4.4.5 Data analysis

Micorarray data of each slide were loaded into the NOMAD database and normalized using the default settings. In brief, a scalar normalization factor was calculated for each array using unfiltered high quality features with background subtracted median intensities greater than zero for each channel and a pixel regression correlation coefficient greater than or equal to 0.75. The data of all slides were extracted from the database and log-transformed for further analysis by filtering the spots of poor quality, flagged, or spots for which intensities in two channels were close to background (median of intensity less than 2 median of background plus 2 standard deviations for both Cy3 and Cy5 signals). To analyze the perturbations of gene expression under drug treatment, each gene profile of

drug treatment was subtracted by its negative control. Differential expression genes were extracted by a double three-two method. One gene is retained when the peak of change through the time course is larger than 3-fold and at least one neighbor point of the peak has 2-fold change. Clustering analysis was performed by Cluster program and hierarchical tree was viewed by TreeView (Eisen, et al., 1998).

Chatper 5 Computers and databases

Bioinformatics involves the use of mathematical tools to extract, organize and analyze the huge amounts of data produced by high-throughput biological techniques and to solve biological problems usually on the molecular level. Major research efforts in this field include sequence analysis, gene prediction and annotation, genome assembly, protein structure analysis and prediction, prediction of gene expression, protein-protein interactions, and the modeling of evolution. The Use of the program OligoRankPick for oligonucleotide selection (Hu, et al., 2007), analysis of genomic context including phylogenetic profiles and domain predictions for predicting functional linkages (Date and Marcotte, 2003; Lee, et al., 2006), storage and analysis of gene expression profiles, network reconstruction and analysis, and network-based predictions of gene function, require computer infrastructure with adequate processing power and data storage facilities. Such requirements assume even more importance, if real-time, large-scale analysis is planned.

Our goal is to establish a computer network that would allow us to collect and store functional genomic data, such as complete genome sequence data and gene expression data, and use these data to create databases for functional genomic research on *P. falciparum* and also provide the web service for the users, such as OligoRankPick, gene functional network, network-based gene annotations.

5.1 A computer network

To establish a computer network for collecting, storing data, web service and computing task, we establish a "triangular" computer network (figure 5.1). This network is comprised of two independent computers and one cluster of supercomputer. The first computer (ZBLab) is web server for open users which provides web tools to show our results such as oligo information, microarray data, and network. The second computer (S3E3) is used for data storage and small computing task. We linked two computers to the cluster of supercomputer (NTU-SBS-Cluster) which could provide large computing power for large-scale analysis, for example, performing the BLAST searches in the oligonucleotide selection of OligoRankPick. Establishing this infrastructure provides a viable model for a relatively small research group to carry out complex biological analyses that involve large datasets with the genome wide approaches.

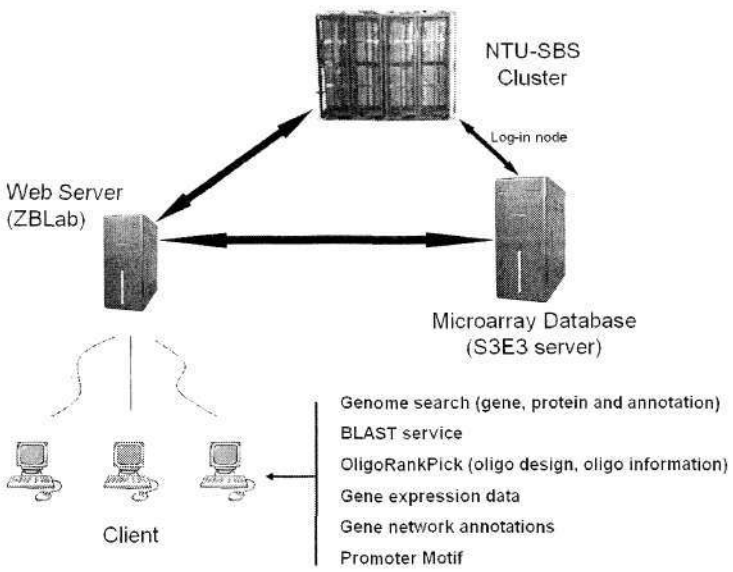


Figure 5.1 The architecture of a computer network for data storage, computing and web service.

5.2 Web service

The network provides several web tools to show the biological information of *P. falciparum* (figure 5.1). Web service is provided to search the probe information for DNA microarrays of malaria species (http://zblab.sbs.ntu.edu/software.html) including the developing address of OligoRankPick program. We also provided the gene, protein and genome information of *P. falciparum* synchronized with PlasmoDB (http://www.plasmodb.org) and the NCBI BLAST online service. Another web searching is the reconstructed gene functional network providing the functional linkages between genes and the new functional predictions based on this network (http://zblab.sbs.ntu.edu.sg/network/). For each gene, different types of functional information are provided and evaluated. Of course, this gene is linked to other databases, such as gene/protein information, KEGG, Gene Ontology and our microarray database of gene expression profiles. Here we try to offer an interface for the users to find the biological knowledge of genes or proteins in *P. falciparum* easily about their sequence, trancriptome, regulation, interactome, molecular function, and cellular process.

5.3 Update databases

The relational databases were built based on MySQL or file system. Perl scripts were developed to update the system. Data source will be automatically downloaded, parsed and used to update the local databases and files.

Chapter 6 Final Summary and Perspective

In this thesis, I developed OligoRankPick which provides a powerful alternative for long oligonucleotide microarray design for genomes with extreme GC content fluctuations and high abundance of highly homologous gene families. In its simplest implementation a user needs only to define the probe length and an expected GC content or Tm. Using this method we have designed high quality and long oligonucleotide DNA microarrays for the parasitic species including *P. falciparum*, *P. vivax* and three rodent malaria parasite *P. chabaudi*, *P. yoelii*, and *P. berghei*. Based on the designed *P. falciparum* DNA microarray, we carried out extensive transcriptional profiling of *P. falciparum* responses to chemically induced growth perturbations and used these data to reconstruct a gene interactome network that allow us to predict function of 2547 *Plasmodium* hypothetical genes. The accuracy of these predictions was demonstrated by the functional validation of 21 selected protein candidates from which 19 localized to intracellular compartments associated with the invasion machinery. These data also demonstrate that transcriptional profiling of growth perturbations provides a powerful technique for functional genomics of *Plasmodium* parasites. Finally, I studied the effects of several classes of inhibitors of protein phosphorylation and dephosphorylation pathways on the growth of *Plasmodium* cell and the regulation of gene expression. The growth assays showed that these inhibitors effectively inhibit the erythrocyte invasion by *P. falciparum in vitro*. The global gene expression profiling shows that

diverse but specific transcriptional responses induced by these inhibitors were observed when parasite cells were treated at the early schizont stage. Interestingly, several transcription factors and signaling genes were significantly regulated resulting from the inhibition of calcium dependent signaling and calcineurin signaling pathways, which suggests that the phosphorylation and/or dephosphorylation play vital roles in the gene expression regulation in *P. falciparum*. All these microarray information, gene expression data, network, gene predictions and other databases and developed bioinformatics tools are available online on the following website: http://zblab.sbs.ntu.edu.sg/program.html.

Although this network covers 88% genome of *P. falciparum* and about 2547 hypothetical proteins were assigned functional clues, to improve the accuracy of the network and predictions of gene function, more data will be incorporated in the future. For example, more growth perturbation data and proteomic expression profiles are being investigated. Currently, 22 proteins were chosen for intracellular localization analysis to evaluate the network-based predictions of gene function. These invasion proteins would be further analyzed to illustrate the molecular function in the invasion process in the future, which would promote deep understanding of the mechanism of merozoite invasion in *P. falciparum*. Furthermore, more proteins would be selected for function analysis, such as proteins associated histone modification and DNA binding proteins. I also am trying to identify protein changes in the perturbations of kinase inhibitors using the 2D-DIGE/MS proteomic technique. We hope to establish the putative regulation

network between protein kinases, transcription factors and responded genes, as well as with the knowledge of gene function network, to illustrate the mode-of-actions of these inhibitors and their roles in the cell cycle progression in *P. falciparum*.

References

Aikawa, M., Miller, L.H., Johnson, J. and Rabbege, J. (1978) Erythrocyte entry by malarial parasites. A moving junction between erythrocyte and parasite, *J Cell Biol*, **77**, 72-82.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Arastu-Kapur, S., Ponder, E.L., Fonovic, U.P., Yeoh, S., Yuan, F., Fonovic, M., Grainger, M., Phillips, C.I., Powers, J.C. and Bogyo, M. (2008) Identification of proteases that regulate erythrocyte rupture by the malaria parasite Plasmodium falciparum, *Nat Chem Biol*, **4**, 203-213.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Balu, B., Shoue, D.A., Fraser, M.J., Jr. and Adams, J.H. (2005) High-efficiency transformation of Plasmodium falciparum by the lepidopteran transposable element piggyBac, *Proc Natl Acad Sci U S A*, **102**, 16391-16396.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet*, **5**, 101-113.

Barnwell, J.W. and Galinski, M.R. (1998) *Invasion of vertebrate cells: erythrocytes. In Sherman,I.W. (ed.), Malaria: Parasite Biology, Pathogenesis and Protection. ASM Press, Washington, DC, pp. 193.*

Baum, J., Richard, D., Healer, J., Rug, M., Krnajski, Z., Gilberger, T.W., Green, J.L., Holder, A.A. and Cowman, A.F. (2006) A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites, *J Biol Chem*, **281**, 5197-5208.

Baum, J., Tonkin, C.J., Paul, A.S., Rug, M., Smith, B.J., Gould, S.B., Richard, D.,

Pollard, T.D. and Cowman, A.F. (2008) A malaria parasite formin regulates actin polymerization and localizes to the parasite-erythrocyte moving junction during invasion, *Cell Host Microbe*, **3**, 188-198.

Bell, A., Monaghan, P. and Page, A.P. (2006) Peptidyl-prolyl cis-trans isomerases (immunophilins) and their roles in parasite biochemistry, host-parasite interaction and antiparasitic drug action, *Int J Parasitol*, **36**, 261-276.

Bell, A., Wernli, B. and Franklin, R.M. (1994) Roles of peptidyl-prolyl cis-trans isomerase and calcineurin in the mechanisms of antimalarial action of cyclosporin A, FK506, and rapamycin, *Biochem Pharmacol*, **48**, 495-503.

Bhardwaj, N. and Lu, H. (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes, *Bioinformatics*, **21**, 2730-2738.

Billker, O., Dechamps, S., Tewari, R., Wenig, G., Franke-Fayard, B. and Brinkmann, V. (2004) Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite, *Cell*, **117**, 503-514.

Birkholtz, L.M., Bastien, O., Wells, G., Grando, D., Joubert, F., Kasam, V., Zimmermann, M., Ortet, P., Jacq, N., Saidani, N., Roy, S., Hofmann-Apitius, M., Breton, V., Louw, A.I. and Marechal, E. (2006) Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space?, *Malar J*, **5**, 110.

Blackman, M.J. (2000) Proteases involved in erythrocyte invasion by the malaria parasite: function and potential as chemotherapeutic targets, *Curr Drug Targets*, **1**, 59-83.

Blackman, M.J. and Bannister, L.H. (2001) Apical organelles of Apicomplexa: biology and isolation by subcellular fractionation, *Mol Biochem Parasitol*, **117**, 11-25.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford, D.K., Melton, D.A., Jaenisch, R. and Young, R.A. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells, *Cell*, **122**, 947-956.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J. and DeRisi, J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum, *PLoS Biol*, **1**, E5.

Bozdech, Z., Mok, S., Hu, G., Imwong, M., Russell, B., Ginsburg, H., Carlton, J., White, N. and Preiser, P. (2008) Transcriptome of Plasmodium vivax: the complex evolution of stage specific transcriptional regulation in malaria parasite species, *submitted*.

Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B. and DeRisi, J.L. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray, *Genome Biol*, **4**, R9.

Brohee, S. and van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks, *BMC Bioinformatics*, **7**, 488.

Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays, *Nat Genet*, **21**, 33-37.

Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K., Miller-Graziano, C., Moldawer, L.L., Mindrinos, M.N., Davis, R.W., Tompkins, R.G. and Lowry, S.F. (2005) A network-based analysis of systemic inflammation in humans, *Nature*, **437**, 1032-1037.

Canduri, F., Perez, P.C., Caceres, R.A. and de Azevedo, W.F., Jr. (2007) Protein kinases as targets for antiparasitic chemotherapy drugs, *Curr Drug Targets*, **8**, 389-398.

Carlton, J. (2003) The Plasmodium vivax genome sequencing project, *Trends Parasitol*, **19**, 227-231.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E.,

Harris, M.A., Cunningham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J. and Carucci, D.J. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii, *Nature*, **419**, 512-519.

Carruthers, V.B. and Sibley, L.D. (1999) Mobilization of intracellular calcium stimulates microneme discharge in Toxoplasma gondii, *Mol Microbiol*, **31**, 421-428.

Carter, M.G., Sharov, A.A., VanBuren, V., Dudekula, D.B., Carmack, C.E., Nelson, C. and Ko, M.S. (2005) Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray, *Genome Biol*, **6**, R61.

Chattopadhyay, R., Rathore, D., Fujioka, H., Kumar, S., de la Vega, P., Haynes, D., Moch, K., Fryauff, D., Wang, R., Carucci, D.J. and Hoffman, S.L. (2003) PfSPATR, a Plasmodium falciparum protein containing an altered thrombospondin type I repeat domain is expressed at several stages of the parasite life cycle and is the target of inhibitory antibodies, *J Biol Chem*, **278**, 25977-25981.

Chitnis, C.E. and Blackman, M.J. (2000) Host cell invasion by malaria parasites, *Parasitol Today*, **16**, 411-415.

Choubey, V., Maity, P., Guha, M., Kumar, S., Srivastava, K., Puri, S.K. and Bandyopadhyay, U. (2007) Inhibition of Plasmodium falciparum choline kinase by hexadecyltrimethylammonium bromide: a possible antimalarial mechanism, *Antimicrob Agents Chemother*, **51**, 696-706.

Chua, H.N., Sung, W.K. and Wong, L. (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions, *Bioinformatics*, **22**, 1623-1630.

Coombs, G.H., Goldberg, D.E., Klemba, M., Berry, C., Kay, J. and Mottram, J.C. (2001) Aspartic proteases of Plasmodium falciparum and other parasitic protozoa as drug targets, *Trends Parasitol*, **17**, 532-537.

Coulson, R.M., Hall, N. and Ouzounis, C.A. (2004) Comparative genomics of

transcriptional control in the human malaria parasite Plasmodium falciparum, *Genome Res*, **14**, 1548-1554.

Cowman, A.F. and Crabb, B.S. (2002) The Plasmodium falciparum genome--a blueprint for erythrocyte invasion, *Science*, **298**, 126-128.

Cowman, A.F. and Crabb, B.S. (2006) Invasion of red blood cells by malaria parasites, *Cell*, **124**, 755-766.

Dahl, E.L., Shock, J.L., Shenai, B.R., Gut, J., DeRisi, J.L. and Rosenthal, P.J. (2006) Tetracyclines specifically target the apicoplast of the malaria parasite Plasmodium falciparum, *Antimicrob Agents Chemother*, **50**, 3124-3131.

Date, S.V. and Marcotte, E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages, *Nat Biotechnol*, **21**, 1055-1062.

Date, S.V. and Stoeckert, C.J., Jr. (2006) Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale, *Genome Res*, **16**, 542-549.

David, F.N. (1949) The moments of the z and F distributions, *Biometrika*, **36**, 394-403.

de Hoog, C.L. and Mann, M. (2004) Proteomics, *Annu Rev Genomics Hum Genet*, **5**, 267-293.

de Lanerolle, P., Johnson, T. and Hofmann, W.A. (2005) Actin and myosin I in the nucleus: what next?, *Nat Struct Mol Biol*, **12**, 742-746.

De Silva, E.K., Gehrke, A.R., Olszewski, K., Leon, I., Chahal, J.S., Bulyk, M.L. and Llinas, M. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors, *Proc Natl Acad Sci U S A*, **105**, 8393-8398.

Deng, M., Mehta, S., Sun, F. and Chen, T. (2002) Inferring domain-domain interactions from protein-protein interactions, *Genome Res*, **12**, 1540-1548.

Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. (2003) Prediction of protein function using protein-protein interaction data, *J Comput Biol*, **10**, 947-960.

DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680-686.

Dhalluin, C., Carlson, J.E., Zeng, L., He, C., Aggarwal, A.K. and Zhou, M.M. (1999)

Structure and ligand of a histone acetyltransferase bromodomain, *Nature*, **399**, 491-496.

Dluzewski, A.R. and Garcia, C.R. (1996) Inhibition of invasion and intraerythrocytic development of Plasmodium falciparum by kinase inhibitors, *Experientia*, **52**, 621-623.

Dobson, S., May, T., Berriman, M., Del Vecchio, C., Fairlamb, A.H., Chakrabarti, D. and Barik, S. (1999) Characterization of protein Ser/Thr phosphatases of the malaria parasite, Plasmodium falciparum: inhibition of the parasitic calcineurin by cyclophilin-cyclosporin complex, *Mol Biochem Parasitol*, **99**, 167-181.

Doerig, C., Endicott, J. and Chakrabarti, D. (2002) Cyclin-dependent kinase homologues of Plasmodium falciparum, *Int J Parasitol*, **32**, 1575-1585.

Doerig, C. and Meijer, L. (2007) Antimalarial drug discovery: targeting protein kinases, *Expert Opin Ther Targets*, **11**, 279-290.

Dowse, T. and Soldati, D. (2004) Host cell invasion by the apicomplexans: the significance of microneme protein proteolysis, *Curr Opin Microbiol*, **7**, 388-396.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, **95**, 14863-14868.

Ellermeier, C. and Smith, G.R. (2005) Cohesins are required for meiotic DNA breakage and recombination in Schizosaccharomyces pombe, *Proc Natl Acad Sci U S A*, **102**, 10952-10957.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events, *Nature*, **402**, 86-90.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res*, **30**, 1575-1584.

Fidock, D.A. and Wellems, T.E. (1997) Transformation with human dihydrofolate reductase renders malaria parasites insensitive to WR99210 but does not affect the intrinsic activity of proguanil, *Proc Natl Acad Sci U S A*, **94**, 10931-10936.

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D.,

Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R. and Carucci, D.J. (2002) A proteomic view of the Plasmodium falciparum life cycle, *Nature*, **419**, 520-526.

Gantt, S., Persson, C., Rose, K., Birkett, A.J., Abagyan, R. and Nussenzweig, V. (2000) Antibodies against thrombospondin-related anonymous protein do not inhibit Plasmodium sporozoite infectivity in vivo, *Infect Immun*, **68**, 3667-3673.

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

Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae, *Nat Genet*, **29**, 482-486.

Ginsburg, H. (2008) Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium, *Trends Parasitol*.

Glockner, G., Eichinger, L., Szafranski, K., Pachebat, J.A., Bankier, A.T., Dear, P.H., Lehmann, R., Baumgart, C., Parra, G., Abril, J.F., Guigo, R., Kumpf, K., Tunggal, B., Cox, E., Quail, M.A., Platzer, M., Rosenthal, A. and Noegel, A.A. (2002) Sequence and analysis of chromosome 2 of Dictyostelium discoideum, *Nature*, **418**, 79-85.

Groth, P., Weiss, B., Pohlenz, H.D. and Leser, U. (2008) Mining phenotypes for gene function prediction, *BMC Bioinformatics*, **9**, 136.

Gunasekera, A.M., Myrick, A., Le Roch, K., Winzeler, E. and Wirth, D.F. (2007) Plasmodium falciparum: genome wide perturbations in transcript profiles among

mixed stage cultures after chloroquine treatment, *Exp Parasitol*, **117**, 87-92.

Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D. and Wang, J. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network, *Bioinformatics*, **23**, 2121-2128.

Haase, S., Cabrera, A., Langer, C., Treeck, M., Struck, N., Herrmann, S., Jansen, P.W., Bruchhaus, I., Bachmann, A., Dias, S., Cowman, A.F., Stunnenberg, H.G., Spielmann, T. and Gilberger, T.W. (2008) Characterization of a conserved rhoptry-associated leucine zipper-like protein in the malaria parasite Plasmodium falciparum, *Infect Immun*, **76**, 879-887.

Harrison, P.M. and Gerstein, M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution, *J Mol Biol*, **318**, 1155-1174.

Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology, *Nature*, **402**, C47-52.

Hay, S.I., Guerra, C.A., Tatem, A.J., Noor, A.M. and Snow, R.W. (2004) The global distribution and population at risk of malaria: past, present, and future, *Lancet Infect Dis*, **4**, 327-336.

Hayes, J.J. and Hansen, J.C. (2002) New insights into unwrapping DNA from the nucleosome from a single-molecule optical tweezers method, *Proc Natl Acad Sci U S A*, **99**, 1752-1754.

He, Z., Wu, L., Li, X., Fields, M.W. and Zhou, J. (2005) Empirical establishment of oligonucleotide probe design criteria, *Appl Environ Microbiol*, **71**, 3753-3760.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) Assessment of prediction accuracy of protein function from protein--protein interaction data, *Yeast*, **18**, 523-531.

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

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis,

C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H. and Linsley, P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer, *Nat Biotechnol*, **19**, 342-347.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000) Functional discovery via a compendium of expression profiles, *Cell*, **102**, 109-126.

Huttenhower, C., Hibbs, M., Myers, C. and Troyanskaya, O.G. (2006) A scalable method for integration and functional analysis of multiple microarray datasets, *Bioinformatics*, **22**, 2890-2897.

Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S.M., Bianchettin, G., Borzym, K., Bothe, G., Bruschi, C.V., Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R.M., Cronin, A., Cruz, A.K., Davies, R.M., De Gaudenzi, J., Dobson, D.E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A.C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J.C., Muller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O'Neil, S., Pentony, M., Pohl, T.M., Price, C., Purnelle, B., Quail, M.A., Rabbinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J.C., Rutter, S., Saunders, D., Schafer, M., Schein, J., Schwartz, D.C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D.F., Blackwell, J.M., Stuart, K.D., Barrell, B. and Myler, P.J. (2005) The genome of the kinetoplastid parasite, Leishmania major, *Science*, **309**, 436-442.

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, *Nature*, **409**, 533-538.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, **302**, 449-453.

Jeanmougin, F., Wurtz, J.M., Le Douarin, B., Chambon, P. and Losson, R. (1997) The bromodomain revisited, *Trends Biochem Sci*, **22**, 151-153.

Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays, *Nucleic Acids Res*, **28**, 4552-4557.

Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002) The KEGG databases at GenomeNet, *Nucleic Acids Res*, **30**, 42-46.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome, *Nucleic Acids Res*, **32**, D277-280.

Kappes, B., Doerig, C.D. and Graeser, R. (1999) An overview of Plasmodium protein kinases, *Parasitol Today*, **15**, 449-454.

Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R. and Kasif, S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks, *Proc Natl Acad Sci U S A*, **101**, 2888-2893.

Kato, N., Sakata, T., Breton, G., Le Roch, K.G., Nagle, A., Andersen, C., Bursulaya, B., Henson, K., Johnson, J., Kumar, K.A., Marr, F., Mason, D., McNamara, C., Plouffe, D., Ramachandran, V., Spooner, M., Tuntland, T., Zhou, Y., Peters, E.C., Chatterjee, A., Schultz, P.G., Ward, G.E., Gray, N., Harper, J. and Winzeler, E.A. (2008) Gene expression signatures and small-molecule compounds link a protein kinase to Plasmodium falciparum motility, *Nat Chem Biol*, **4**, 347-356.

Kawamoto, F., Fujioka, H., Murakami, R., Syafruddin, Hagiwara, M., Ishikawa, T. and Hidaka, H. (1993) The roles of Ca2+/calmodulin- and cGMP-dependent pathways in gametogenesis of a rodent malaria parasite, Plasmodium berghei,

Eur J Cell Biol, **60**, 101-107.

Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, F.C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell*, **9**, 1133-1143.

Khan, S.M., Franke-Fayard, B., Mair, G.R., Lasonder, E., Janse, C.J., Mann, M. and Waters, A.P. (2005) Proteome analysis of separated male and female gametocytes reveals novel sex-specific Plasmodium biology. *Cell*, **121**, 675-687.

Khanin, R. and Wit, E. (2007) *"Construction of Malaria Gene Expression Network Using Partial Correlations" in Methods of Microarray Data Analysis V*. Springer, New York.

Kieschnick, H., Wakefield, T., Narducci, C.A. and Beckers, C. (2001) Toxoplasma gondii attachment to host cells is regulated by a calmodulin-like domain protein kinase. *J Biol Chem*, **276**, 12369-12377.

Kim, W.K., Krumpelman, C. and Marcotte, E.M. (2008) Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biol*, **9 Suppl 1**, S5.

Kobe, B. and Deisenhofer, J. (1994) The leucine-rich repeat: a versatile binding motif. *Trends Biochem Sci*, **19**, 415-421.

Kooij, T.W., Janse, C.J. and Waters, A.P. (2006) Plasmodium post-genomics: better the bug you know?. *Nat Rev Microbiol*, **4**, 344-357.

Kotaka, M., Ye, H., Alag, R., Hu, G., Bozdech, Z., Preiser, P.R., Yoon, H.S. and Lescar, J. (2008) Crystal structure of the FK506 binding domain of Plasmodium falciparum FKBP35 in complex with FK506. *Biochemistry*, **47**, 5951-5961.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E.,

Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature*, **440**, 637-643.

Kumar, R., Adams, B., Musiyenko, A., Shulyayeva, O. and Barik, S. (2005) The FK506-binding protein of the malaria parasite, Plasmodium falciparum, is a FK506-sensitive chaperone with FK506-independent calcineurin-inhibitory activity, *Mol Biochem Parasitol*, **141**, 163-173.

Kyes, S., Horrocks, P. and Newbold, C. (2001) Antigenic variation at the infected red cell surface in malaria, *Annu Rev Microbiol*, **55**, 673-707.

LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S. and Hughes, R.E. (2005) A protein interaction network of the malaria parasite Plasmodium falciparum, *Nature*, **438**, 103-107.

Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G. and Mann, M. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry, *Nature*, **419**, 537-542.

Laurent, D., Jullian, V., Parenty, A., Knibiehler, M., Dorin, D., Schmitt, S., Lozach, O., Lebouvier, N., Frostin, M., Alby, F., Maurel, S., Doerig, C., Meijer, L. and Sauvain, M. (2006) Antimalarial potential of xestoquinone, a protein kinase inhibitor isolated from a Vanuatu marine sponge Xestospongia sp, *Bioorg Med Chem*, **14**, 4477-4482.

Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J. and Winzeler, E.A. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, **301**, 1503-1508.

Lee, H., Deng, M., Sun, F. and Chen, T. (2006) An integrated approach to the prediction of domain-domain interactions, *BMC Bioinformatics*, **7**, 269.

Lee, I., Date, S.V., Adai, A.T. and Marcotte, E.M. (2004) A probabilistic functional network of yeast genes, *Science*, **306**, 1555-1558.

Li, F. and Stormo, G.D. (2001) Selection of optimal DNA oligos for gene expression arrays, *Bioinformatics*, **17**, 1067-1076.

Li, X., He, Z. and Zhou, J. (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation, *Nucleic Acids Res*, **33**, 6114-6123.

Liu, J., Farmer, J.D., Jr., Lane, W.S., Friedman, J., Weissman, I. and Schreiber, S.L. (1991) Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes, *Cell*, **66**, 807-815.

Llinas, M., Bozdech, Z., Wong, E.D., Adai, A.T. and DeRisi, J.L. (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains, *Nucleic Acids Res*, **34**, 1166-1173.

Lovett, J.L. and Sibley, L.D. (2003) Intracellular calcium stores in Toxoplasma gondii govern invasion of host cells, *J Cell Sci*, **116**, 3009-3016.

MacCarthy, T., Pomiankowski, A. and Seymour, R. (2005) Using large-scale perturbations in gene network reconstruction, *BMC Bioinformatics*, **6**, 11.

Madeira, L., DeMarco, R., Gazarini, M.L., Verjovski-Almeida, S. and Garcia, C.R. (2003) Human malaria parasites display a receptor for activated C kinase ortholog, *Biochem Biophys Res Commun*, **306**, 995-1001.

Maier, A.G., Rug, M., O'Neill, M.T., Brown, M., Chakravorty, S., Szestak, T., Chesson, J., Wu, Y., Hughes, K., Coppel, R.L., Newbold, C., Beeson, J.G., Craig, A., Crabb, B.S. and Cowman, A.F. (2008) Exported proteins required for virulence and rigidity of Plasmodium falciparum-infected human erythrocytes, *Cell*, **134**, 48-61.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function, *Nature*, **402**, 83-86.

Mathews, C.K., Moen, L.K., Wang, Y. and Sargent, R.G. (1988) Intracellular organization of DNA precursor biosynthetic enzymes, *Trends Biochem Sci*, **13**, 394-397.

Means, A.R. (2000) Regulatory cascades involving calmodulin-dependent protein

kinases, *Mol Endocrinol*, **14**, 4-13.

Mendis, K., Sina, B.J., Marchesini, P. and Carter, R. (2001) The neglected burden of Plasmodium vivax malaria, *Am J Trop Med Hyg*, **64**, 97-106.

Moore, J.T., Silversmith, R.E., Maley, G.F. and Maley, F. (1993) T4-phage deoxycytidylate deaminase is a metalloprotein containing two zinc atoms per subunit, *J Biol Chem*, **268**, 2288-2291.

Morrissette, N.S. and Sibley, L.D. (2002) Disruption of microtubules uncouples budding and nuclear division in Toxoplasma gondii, *J Cell Sci*, **115**, 1017-1025.

Moudy, R., Manning, T.J. and Beckers, C.J. (2001) The loss of cytoplasmic potassium upon host cell breakdown triggers egress of Toxoplasma gondii, *J Biol Chem*, **276**, 41492-41501.

Murali, T.M., Wu, C.J. and Kasif, S. (2006) The art of gene function prediction, *Nat Biotechnol*, **24**, 1474-1475; author reply 1475-1476.

Nielsen, H.B., Wernersson, R. and Knudsen, S. (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays, *Nucleic Acids Res*, **31**, 3491-3496.

Nirmalan, N., Sims, P.F. and Hyde, J.E. (2004) Translational up-regulation of antifolate drug targets in the human malaria parasite Plasmodium falciparum upon challenge with inhibitors, *Mol Biochem Parasitol*, **136**, 63-70.

Nordberg, E.K. (2005) YODA: selecting signature oligonucleotides, *Bioinformatics*, **21**, 1365-1370.

O'Connor, R.J., Schaley, J.E., Feeney, G. and Hearing, P. (2001) The p107 tumor suppressor induces stable E2F DNA binding to repress target promoters, *Oncogene*, **20**, 1882-1891.

O'Donnell, R.A. and Blackman, M.J. (2005) The role of malaria merozoite proteases in red blood cell invasion, *Curr Opin Microbiol*, **8**, 422-427.

O'Donnell, R.A., Freitas-Junior, L.H., Preiser, P.R., Williamson, D.H., Duraisingh, M., McElwain, T.F., Scherf, A., Cowman, A.F. and Crabb, B.S. (2002) A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of Plasmodium falciparum chromosomes, *Embo J*, **21**, 1231-1239.

Oakley, M.S., Kumar, S., Anantharaman, V., Zheng, H., Mahajan, B., Haynes, J.D., Moch, J.K., Fairhurst, R., McCutchan, T.F. and Aravind, L. (2007) Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic Plasmodium falciparum parasites, *Infect Immun*, **75**, 2012-2025.

Obmolova, G., Ban, C., Hsieh, P. and Yang, W. (2000) Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA, *Nature*, **407**, 703-710.

Ono, T., Cabrita-Santos, L., Leitao, R., Bettiol, E., Purcell, L.A., Diaz-Pulido, O., Andrews, L.B., Tadakuma, T., Bhanot, P., Mota, M.M. and Rodriguez, A. (2008) Adenylyl cyclase alpha and cAMP signaling mediate Plasmodium sporozoite apical regulated exocytosis and hepatocyte infection, *PLoS Pathog*, **4**, e1000008.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T., Churcher, C.M., Bentley, S.D., Mungall, K.L., Cerdeno-Tarraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G. and Maskell, D.J. (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica, *Nat Genet*, **35**, 32-40.

Pena-Castillo, L., Tasan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K., Krumpelman, C., Tian, W., Obozinski, G., Qi, Y., Mostafavi, S., Lin, G.N., Berriz, G.F., Gibbons, F.D., Lanckriet, G., Qiu, J., Grant, C., Barutcuoglu, Z., Hill, D.P., Warde-Farley, D., Grouios, C., Ray, D., Blake, J.A., Deng, M., Jordan, M.I., Noble, W.S., Morris, Q., Klein-Seetharaman, J., Bar-Joseph, Z., Chen, T., Sun, F., Troyanskaya, O.G., Marcotte, E.M., Xu, D., Hughes, T.R. and Roth, F.P. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence, *Genome Biol*, **9 Suppl 1**,

S2.

Pinder, J., Fowler, R., Bannister, L., Dluzewski, A. and Mitchell, G.H. (2000) Motile systems in malaria merozoites: how is the red blood cell invaded?, *Parasitol Today*, **16**, 240-245.

Ponting, C.P. and Dickens, N.J. (2001) Genome cartography through domain annotation, *Genome Biol*, **2**, Comment 2006.

Preiser, P., Kaviratne, M., Khan, S., Bannister, L. and Jarra, W. (2000) The apical organelles of malaria merozoites: host cell selection, invasion, host immunity and immune evasion, *Microbes Infect*, **2**, 1461-1477.

Price, R.N., Tjitra, E., Guerra, C.A., Yeung, S., White, N.J. and Anstey, N.M. (2007) Vivax malaria: neglected and not benign, *Am J Trop Med Hyg*, **77**, 79-87.

Pujana, M.A., Han, J.D., Starita, L.M., Stevens, K.N., Tewari, M., Ahn, J.S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W.M., Rual, J.F., Levine, D., Rozek, L.S., Gelman, R.S., Gunsalus, K.C., Greenberg, R.A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sole, X., Hernandez, P., Lazaro, C., Nathanson, K.L., Weber, B.L., Cusick, M.E., Hill, D.E., Offit, K., Livingston, D.M., Gruber, S.B., Parvin, J.D. and Vidal, M. (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction, *Nat Genet*, **39**, 1338-1349.

Ralph, S.A., van Dooren, G.G., Waller, R.F., Crawford, M.J., Fraunholz, M.J., Foth, B.J., Tonkin, C.J., Roos, D.S. and McFadden, G.I. (2004) Tropical infectious diseases: metabolic maps and functions of the Plasmodium falciparum apicoplast, *Nat Rev Microbiol*, **2**, 203-216.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002) Hierarchical organization of modularity in metabolic networks, *Science*, **297**, 1551-1555.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins, *Science*, **290**, 2306-2309.

Rengarajan, J., Bloom, B.R. and Rubin, E.J. (2005) Genome-wide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages, *Proc Natl Acad Sci U S A*, **102**, 8327-8332.

Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G. and Fayard, J.M. (2004) ROSO: optimizing oligonucleotide probes for microarrays, *Bioinformatics*, **20**, 271-273.

Riley, R., Lee, C., Sabatti, C. and Eisenberg, D. (2005) Inferring protein domain interactions from databases of interacting proteins, *Genome Biol*, **6**, R89.

Rouillard, J.M., Zuker, M. and Gulari, E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach, *Nucleic Acids Res*, **31**, 3057-3062.

Rowe, J.A. and Kyes, S.A. (2004) The role of Plasmodium falciparum var genes in malaria in pregnancy, *Mol Microbiol*, **53**, 1011-1019.

Sam-Yellowe, T.Y., Shio, H. and Perkins, M.E. (1988) Secretion of Plasmodium falciparum rhoptry protein into the plasma membrane of host erythrocytes, *J Cell Biol*, **106**, 1507-1513.

Sanders, P.R., Gilson, P.R., Cantin, G.T., Greenbaum, D.C., Nebl, T., Carucci, D.J., McConville, M.J., Schofield, L., Hodder, A.N., Yates, J.R., 3rd and Crabb, B.S. (2005) Distinct protein classes including novel merozoite surface antigens in Raft-like membranes of Plasmodium falciparum, *J Biol Chem*, **280**, 40169-40176.

Schlitt, T., Palin, K., Rung, J., Dietmann, S., Lappe, M., Ukkonen, E. and Brazma, A. (2003) From gene networks to gene function, *Genome Res*, **13**, 2568-2576.

Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast, *Nat Biotechnol*, **18**, 1257-1261.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res*, **13**, 2498-2504.

Sharan, R., Ulitsky, I. and Shamir, R. (2007) Network-based prediction of protein

function, *Mol Syst Biol*, **3**, 88.

Silva-Neto, M.A., Atella, G.C. and Shahabuddin, M. (2002) Inhibition of Ca2+/calmodulin-dependent protein kinase blocks morphological differentiation of plasmodium gallinaceum zygotes to ookinetes, *J Biol Chem*, **277**, 14085-14091.

Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y. and Hay, S.I. (2005) The global distribution of clinical episodes of Plasmodium falciparum malaria, *Nature*, **434**, 214-217.

Soldati, D., Foth, B.J. and Cowman, A.F. (2004) Molecular and functional aspects of parasite invasion, *Trends Parasitol*, **20**, 567-574.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains, *Nucleic Acids Res*, **26**, 320-322.

Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction, *J Mol Biol*, **311**, 681-692.

Stringer, J.R. and Keely, S.P. (2001) Genetics of surface antigen expression in Pneumocystis carinii, *Infect Immun*, **69**, 627-639.

Struck, N.S., de Souza Dias, S., Langer, C., Marti, M., Pearce, J.A., Cowman, A.F. and Gilberger, T.W. (2005) Re-defining the Golgi complex in Plasmodium falciparum using the novel Golgi marker PfGRASP, *Journal of Cell Science*, **118**, 5603-5613.

Struck, N.S., de Souza Dias, S., Langer, C., Marti, M., Pearce, J.A., Cowman, A.F. and Gilberger, T.W. (2005) Re-defining the Golgi complex in Plasmodium falciparum using the novel Golgi marker PfGRASP, *J Cell Sci*, **118**, 5603-5613.

Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A., Shi, T. and Li, Y. (2005) Refined phylogenetic profiles method for predicting protein-protein interactions, *Bioinformatics*, **21**, 3409-3415.

Tian, W., Zhang, L.V., Tasan, M., Gibbons, F.D., King, O.D., Park, J., Wunderlich, Z., Cherry, J.M. and Roth, F.P. (2008) Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function, *Genome*

Biol, **9 Suppl 1**, S7.

Tolstrup, N., Nielsen, P.S., Kolberg, J.G., Frankel, A.M., Vissing, H. and Kauppinen, S. (2003) OligoDesign: Optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling, *Nucleic Acids Res*, **31**, 3758-3762.

Tornow, S. and Mewes, H.W. (2003) Functional modules by relating protein interaction networks and gene expression, *Nucleic Acids Res*, **31**, 6283-6289.

Trager, W. and Jensen, J.B. (1997) Continuous culture of Plasmodium falciparum: its impact on malaria research, *Int J Parasitol*, **27**, 989-1006.

Treeck, M., Struck, N.S., Haase, S., Langer, C., Herrmann, S., Healer, J., Cowman, A.F. and Gilberger, T.W. (2006) A conserved region in the EBL proteins is implicated in microneme targeting of the malaria parasite Plasmodium falciparum, *J Biol Chem*, **281**, 31995-32003.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520-525.

Vaid, A., Thomas, D.C. and Sharma, P. (2008) Role of Ca2+/calmodulin-PfPKB signaling pathway in erythrocyte invasion by Plasmodium falciparum, *J Biol Chem*, **283**, 5589-5597.

Valencia, A. and Pazos, F. (2002) Computational methods for the prediction of protein interactions, *Curr Opin Struct Biol*, **12**, 368-373.

van Dooren, G.G., Marti, M., Tonkin, C.J., Stimmler, L.M., Cowman, A.F. and McFadden, G.I. (2005) Development of the endoplasmic reticulum, mitochondrion and apicoplast during the asexual life cycle of Plasmodium falciparum, *Mol Microbiol*, **57**, 405-419.

Waller, R.F., Keeling, P.J., Donald, R.G., Striepen, B., Handman, E., Lang-Unnasch, N., Cowman, A.F., Besra, G.S., Roos, D.S. and McFadden, G.I. (1998) Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum, *Proc Natl Acad Sci U S A*, **95**, 12352-12357.

Waller, R.F., McConville, M.J. and McFadden, G.I. (2004) More plastids in human

parasites?, *Trends Parasitol*, **20**, 54-57.

Wang, S., Nath, N., Fusaro, G. and Chellappan, S. (1999) Rb and prohibitin target distinct regions of E2F1 for repression and respond to different upstream signals, *Mol Cell Biol*, **19**, 7447-7460.

Wang, X. and Seed, B. (2003) Selection of oligonucleotide probes for protein coding sequences, *Bioinformatics*, **19**, 796-802.

Ward, P., Equinet, L., Packer, J. and Doerig, C. (2004) Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote, *BMC Genomics*, **5**, 79.

Westberg, J., Persson, A., Holmberg, A., Goesmann, A., Lundeberg, J., Johansson, K.E., Pettersson, B. and Uhlen, M. (2004) The genome sequence of Mycoplasma mycoides subsp. mycoides SC type strain PG1T, the causative agent of contagious bovine pleuropneumonia (CBPP), *Genome Res*, **14**, 221-227.

Whitehead, K., Kish, A., Pan, M., Kaur, A., Reiss, D.J., King, N., Hohmann, L., DiRuggiero, J. and Baliga, N.S. (2006) An integrated systems approach for understanding cellular responses to gamma radiation, *Mol Syst Biol*, **2**, 47.

Wickramarachchi, T., Devi, Y.S., Mohmmed, A. and Chauhan, V.S. (2008) Identification and characterization of a novel Plasmodium falciparum merozoite apical protein involved in erythrocyte binding and invasion, *PLoS ONE*, **3**, e1732.

Winzeler, E.A. (2006) Applied systems biology and malaria, *Nat Rev Microbiol*, **4**, 145-151.

Wright, M.A. and Church, G.M. (2002) An open-source oligomicroarray standard for human and mouse, *Nat Biotechnol*, **20**, 1082-1083.

Wuchty, S. and Ipsaro, J.J. (2007) A draft of protein interactions in the malaria parasite P. falciparum, *J Proteome Res*, **6**, 1461-1470.

Xiu, M., Kim, J., Sampson, E., Huang, C.Y., Davis, R.J., Paulson, K.E. and Yee, A.S. (2003) The transcriptional repressor HBP1 is a target of the p38 mitogen-activated protein kinase pathway in cell cycle regulation, *Mol Cell Biol*, **23**, 8890-8901.

Yeoh, S., O'Donnell, R.A., Koussis, K., Dluzewski, A.R., Ansell, K.H., Osborne, S.A., Hackett, F., Withers-Martinez, C., Mitchell, G.H., Bannister, L.H., Bryans, J.S., Kettleborough, C.A. and Blackman, M.J. (2007) Subcellular discharge of a serine protease mediates release of invasive malaria parasites from host erythrocytes, *Cell*, **131**, 1072-1083.

Zak, D.E., Pearson, R.K., Vadigepalli, R., Gonye, G.E., Schwaber, J.S. and Doyle, F.J., 3rd (2003) Continuous-time identification of gene expression models, *Omics*, **7**, 373-386.

Zhang, C.G., Gonzales, A.D., Choi, M.W., Chromy, B.A., Fitch, J.P. and McCutchen-Maloney, S.L. (2005) Subcellular proteomic analysis of host-pathogen interactions using human monocytes exposed to Yersinia pestis and Yersinia pseudotuberculosis, *Proteomics*, **5**, 1877-1888.

Zhu, H., Bilgin, M. and Snyder, M. (2003) Proteomics, *Annu Rev Biochem*, **72**, 783-812.

Appendix: supplementary figures and tables

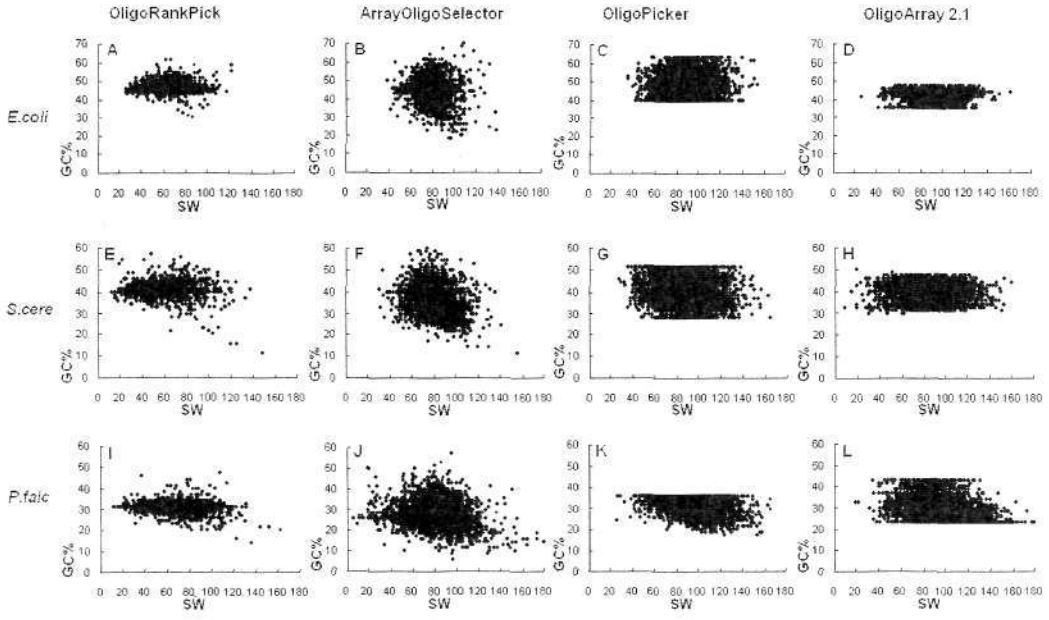


Figure S2.1 The SW (self-binding) score and GC content distributions for the designed oligonucleotide sets. In the each scatter plots SW scores (X-axis) is plotted against the GC content (Y-axis) for all oligonucleotides in the set. Total 12 oligonucleotide sets were designed for three genomes *E. coli* (A-D), *S. cerevisiae* (E-H), and *P. falciparum* (I-L) using all programs OligoRankPick (A, E, I), ArrayOligoSelector (B, F, J), OligoPicker(C, G, K), and OligoArray 2.1 (D, H, L). Tighter distribution the SW scores indicates the improved performance of OligoRankPick for microarray design.

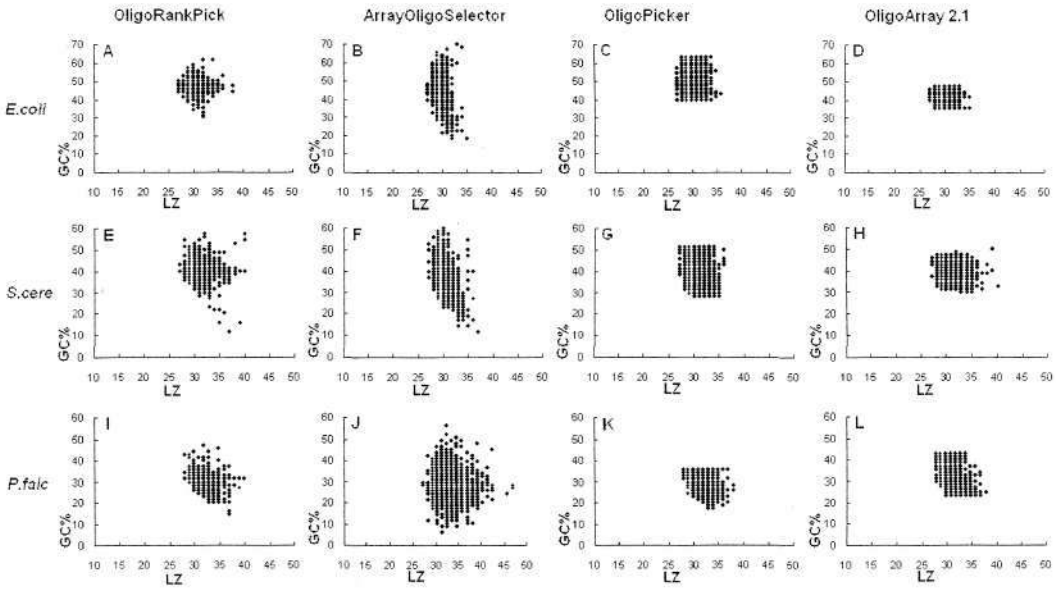


Figure S2.2 The LZ (sequence complexity) score and GC content distributions for the designed oligonucleotide sets. In the each scatter plots the LZ scores (X-axis) are plotted against the GC content (Y-axis) for all oligonucleotides in the set. Total 12 oligonucleotide sets were designed for three genomes *E. coli* (A-D), *S. cerevisiae* (E-H), and *P. falciparum* (I-L) using all programs OligoRankPick (A, E, I), ArrayOligoSelector (B, F, J), OligoPicker(C, G, K), and OligoArray 2.1 (D, H, L). Tighter distribution the LZ scores indicates the improved performance of OligoRankPick for microarray design.

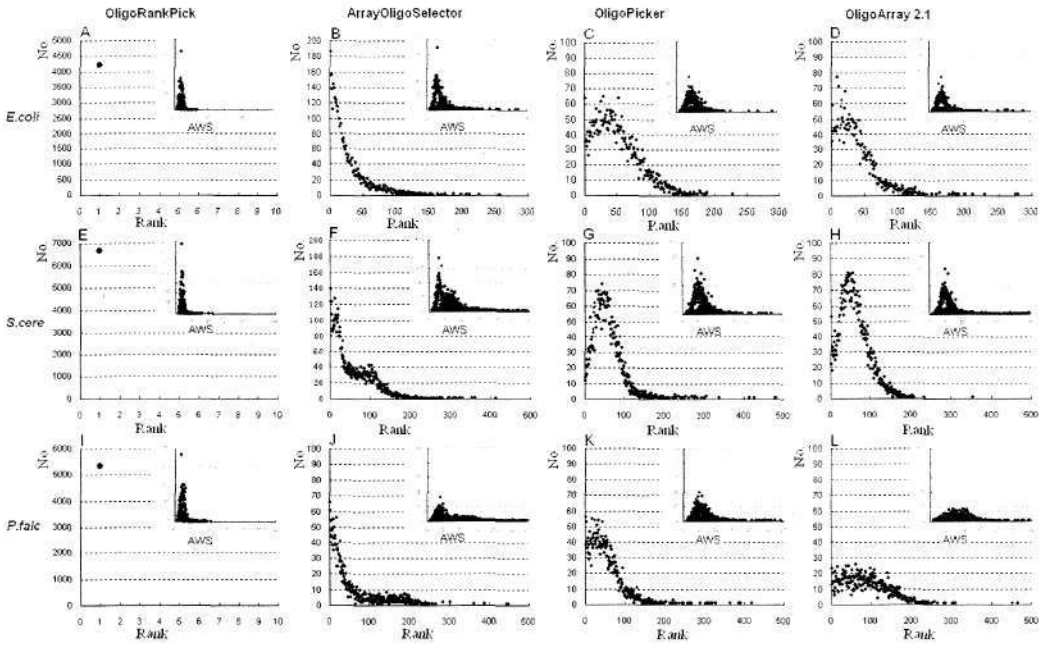


Figure S2.3 The distribution of Rank status and Average weight score of the selected oligonucleotides from three datasets by different programs. A-D, the oligonucleotide sets of *E. coli* by four programs; E-H, the oligonucleotide sets of *S. cerevisiae*; I-L, the oligonucleotide sets of *P. falciparum*. In each diagram, the top-right small diagram is the distribution of AWS (average weight score) of the whole oligonucleotide set, and the diagram below is the rank distribution of the whole oligonucleotide set. For AWS and rank status, the weight set I was first determined by ORP, then all oligonucleotide AWSs were calculated by formula 2 and ranked. The AWS and rank status of all selected oligonucleotides were mapped.

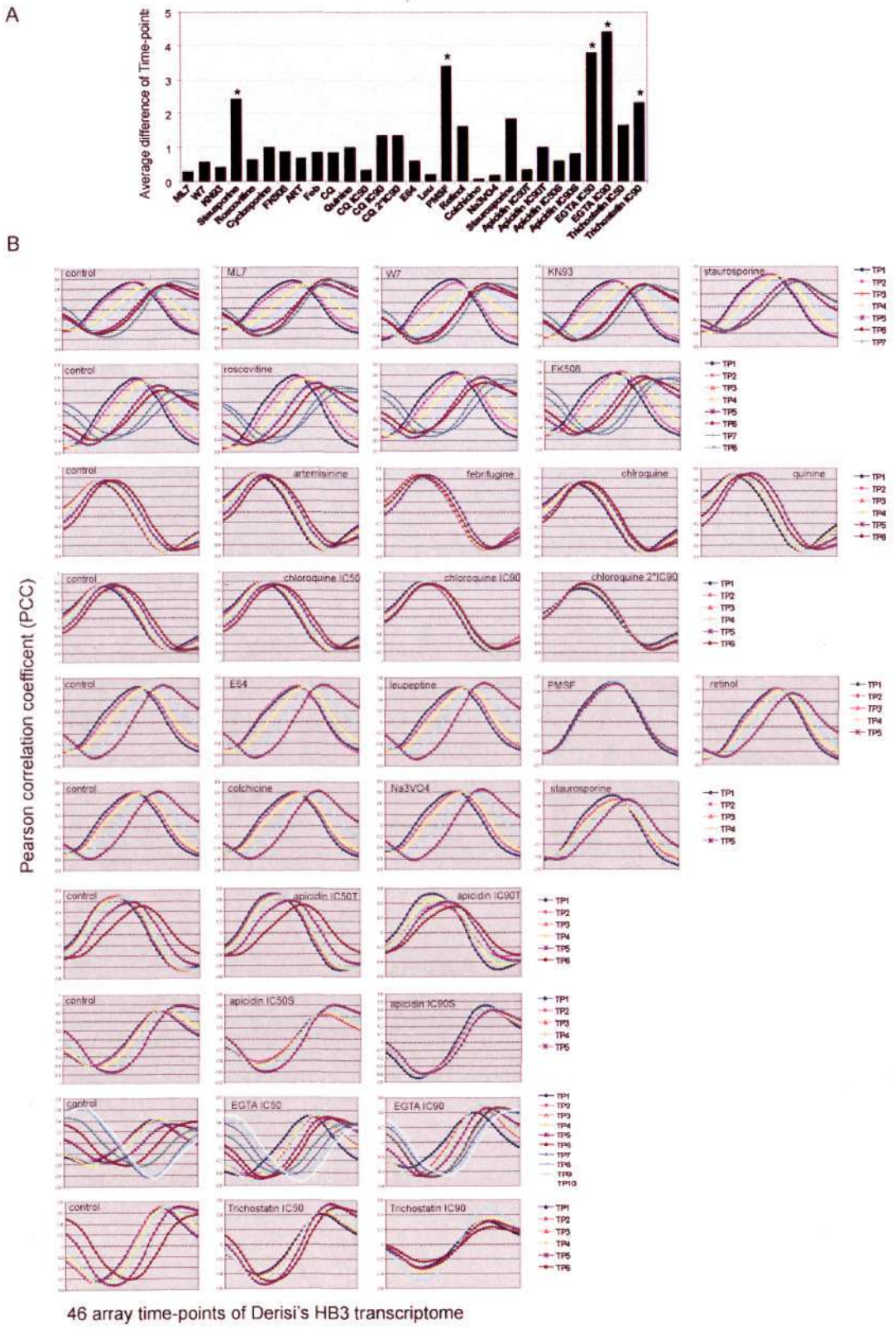


Figure S3.1 Evaluation of growth arrest the during perturbation analyses. To evaluate whether the perturbation induced mRNA profiles correspond to specific responses or to a generic arrest of the IDC transcriptional cascade, we calculate Pearson Correlation Coefficients (PCC) between the microarray results for each

time point in the perturbation time course and the IDC transcriptome (46 microarray data of the HB3 strain IDC transcriptome, http://malaria.ucsf.edu). **B.** For each time point in the perturbation time course (TP1, 2, ...), a function of 46 PCC was assembled for all inhibitor treatments as well as their corresponding untreated controls Table S1. (in the B panel each row of graphs represents one set of experiments and the corresponding control time course is extreme left in each row). The maximum PCC value in each TP profile that corresponds to the best fit a IDC transcriptome the developmental stage of the cells in the perturbation studies (peak PCC time, PPCCT). In each control (untreated cells) time course, the peak PPCCT corresponded well to the expected progression of the parasites cells through the IDC. Thus comparing the PPCCTs between the treatments and controls allows detecting a potential growth/developmental arrest. For this we calculate an average distance between the PPCCT in the perturbation time courses and the corresponding controls **(A.)**. For staurosporine, PMSF, TrichostatinA (IC90) and EGTA, we observe considerably high values of the average time point distances. These high values signal dramatic shifts in the developmental stage. Visual inspection of the PCC profiles confirms a growth arrests that is characterized by retention of the initial (start of the treatment) PPCCT in the treatments while in controls the progression of the PCCT values follow the expected trend (panel B). For all other treatments the PPCCT shifts were considerably smaller (<1.0) which indicate that the mRNA profiles in these treatments do not represent growth arrests but rather correspond to specific transcriptional responses of *P. falciparum* to the perturbations.

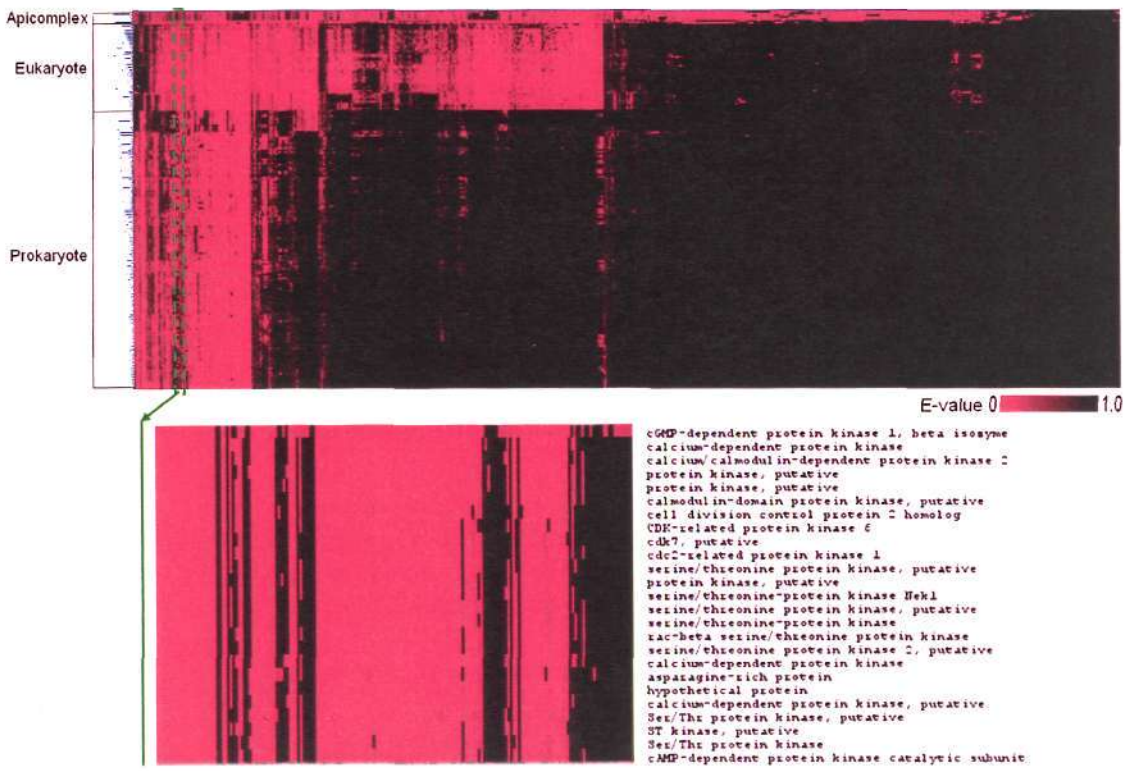


Figure S3.2. Hierarchical clustering of phylogenetic profiles of *P. falciparum*. Organisms varied along the vertical axis, proteins along the horizontal axis. Organisms from the three domains of Apicomplex, eukaryote and prokaryote were separated by black lines on the left. The score of each protein was the top e-value score of the target organism using BLASTP program. Continuous phylogenetic profiles color-coded from red (maximal homology, e-value equal to 0) to black (no homology, e-value equal to 1) A quick look at this figure provides evidence that a lot of evolutionarily meaningful clusters emerged. For example, one cluster of protein kinases (zoomed in figure) suggested that the clusters of proteins with high correlations represent some sort of discrete functional units.

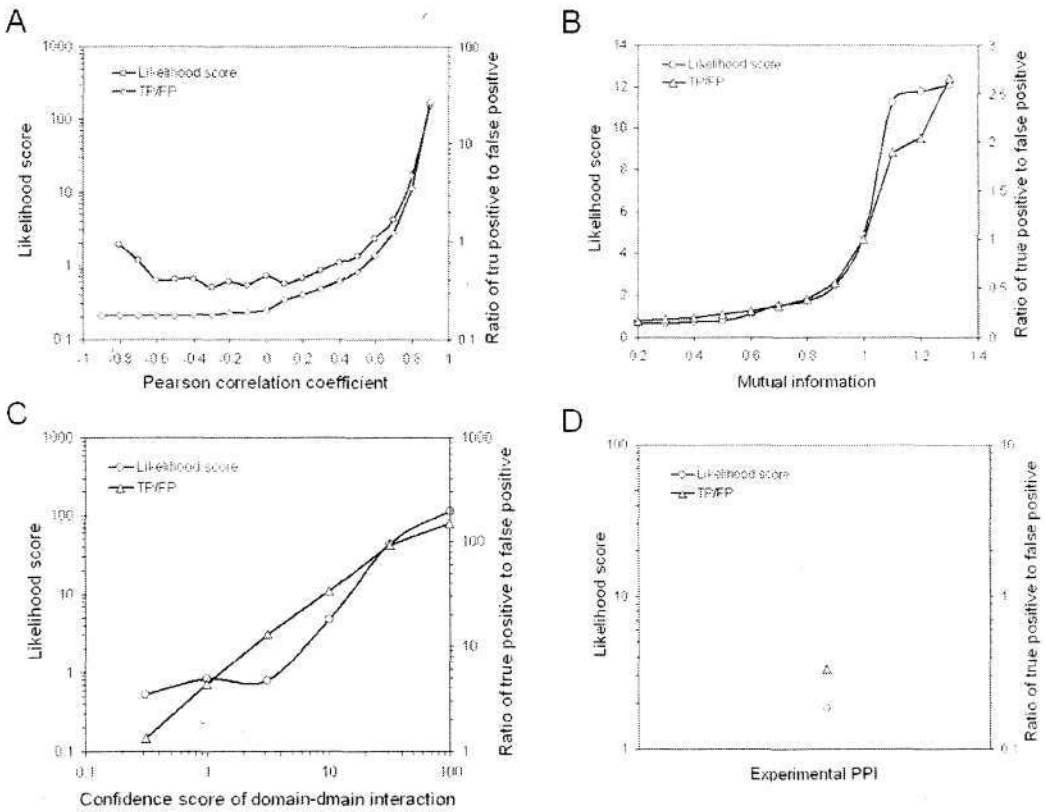


Figure S3.3 Performance of the four input datasets in predicting the gene functional relationships. For each dataset data the likelihood score and the ratio of true positives to false positives based on the benchmark dataset (KEGG pathways) were plotted as function of the binned confidence score. **A.** – the gene expression dataset combining perturbation data and the *IDC* transcriptomes (Llinas, et al., 2006). **B.** – the genomic functional data based on phylogenetic profiles. **C.** – the domain-domain interaction prediction data. **D.** - the protein-protein interaction data determined by yeast two-hybrid system studies. All protein pairs from protein-protein interaction data had the same likelihood score (2.367) and ratio (0.375).

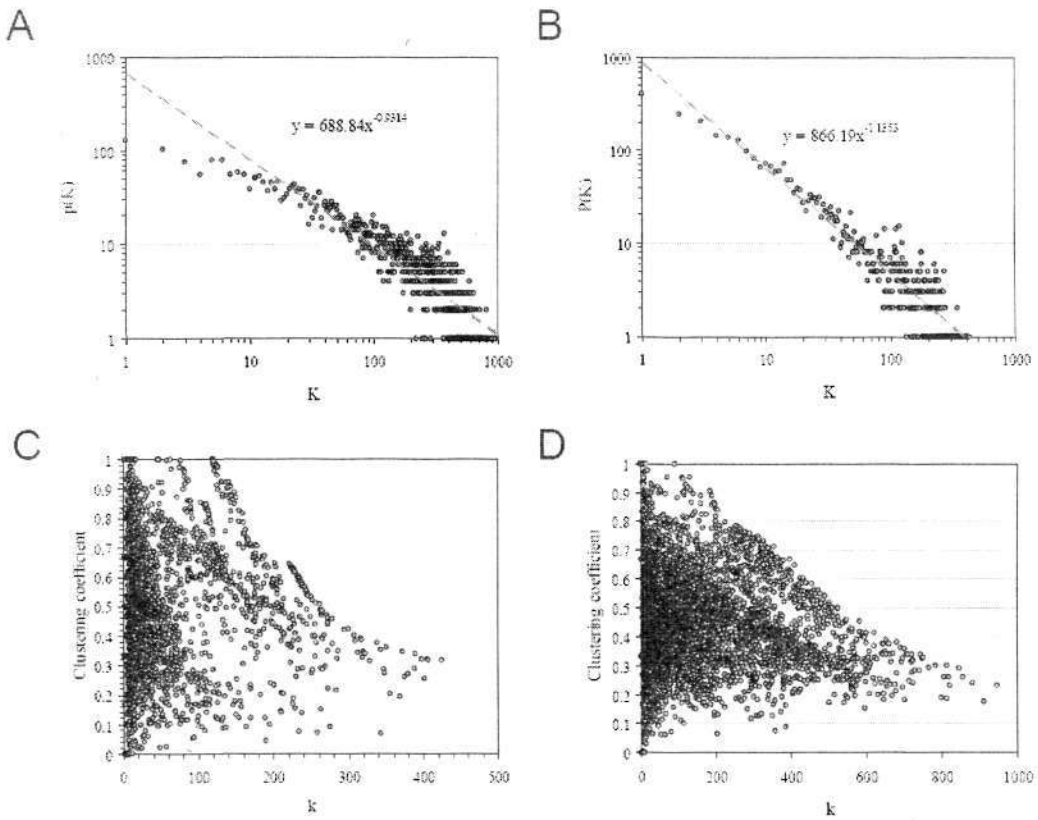
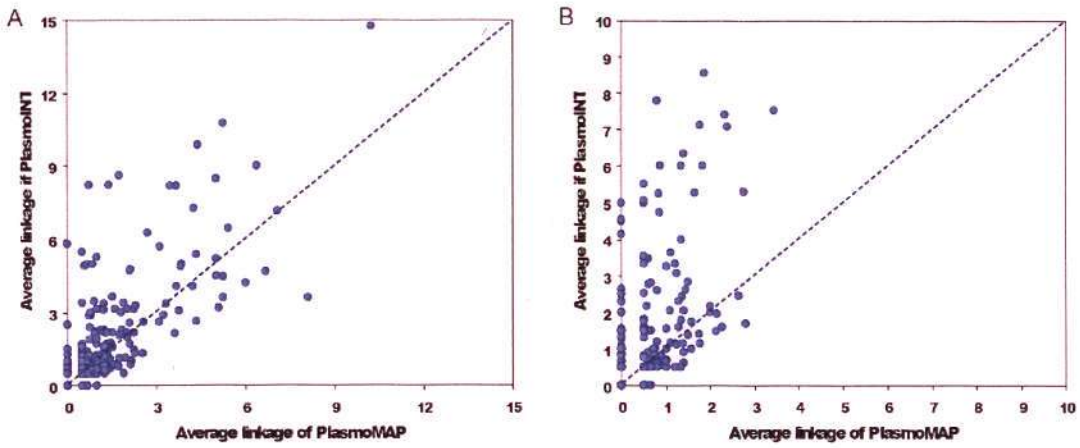


Figure S3.4. Analyses of the network topological structure in the PlasmoINT network. Analysis of the distribution of node connectivity showed that both 50% (A) and 90% (B) confidence level networks had typical scale-free distributions. The distributions of clustering coefficients showed both networks (C and D) were lack of hierarchical structure. The average clustering coefficient was 0.46 and 0.52, respectively.



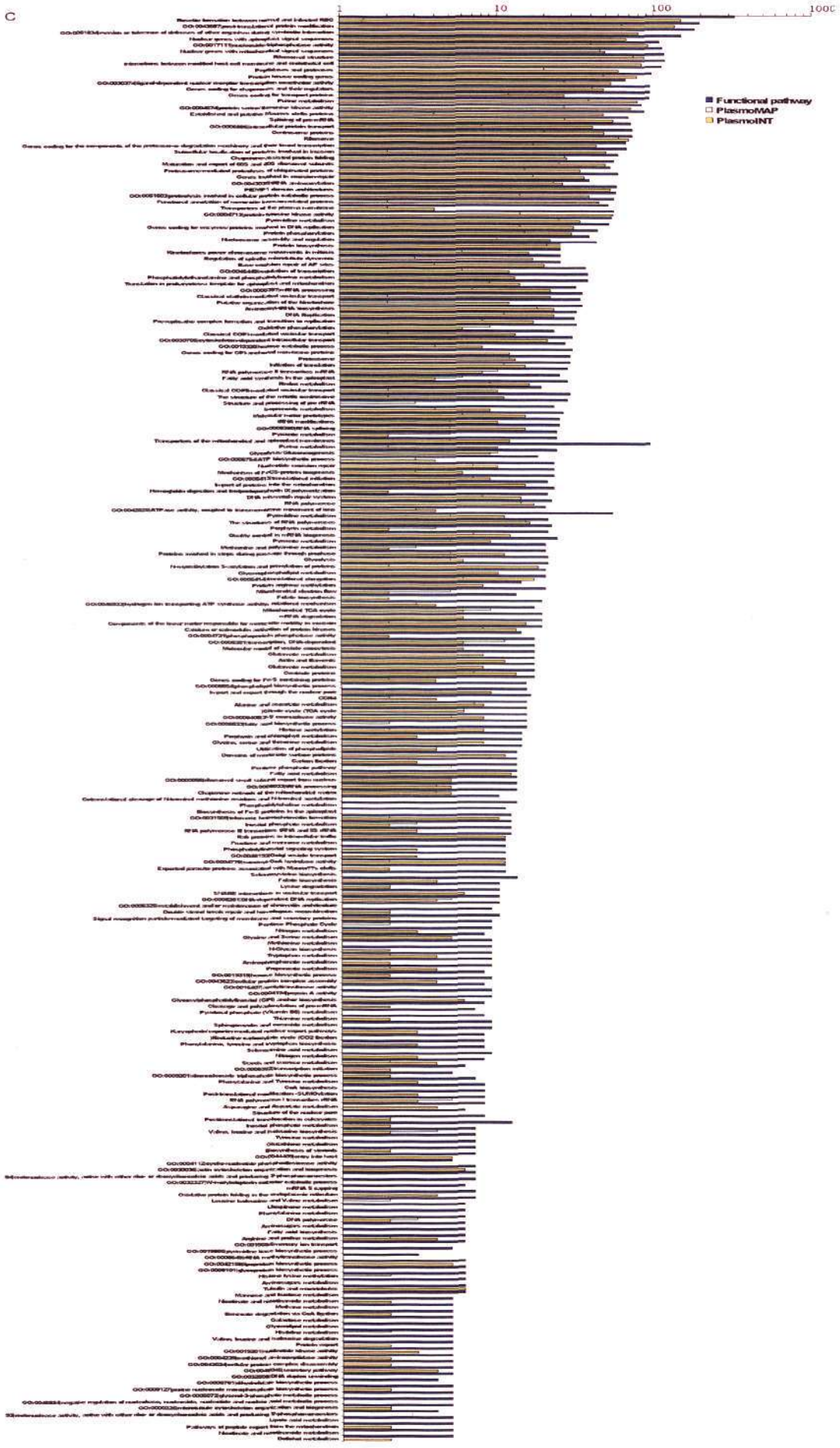


Figure S3.5 Comparisons of the assembled interactome network (PlasmoINT) with a previously reported interactome by Date and Stockert (PlasmoMAP). Comparisons of the average number of linkages in the gene groups corresponding to functional pathways from KEGG, GO and Malaria Parasite Metabolic Pathways (MPMP) were compared for between PlasmoINT and PlasmoMAP for both 50% (**A**) and 90% (**B**) with at least 5 genes present in that pathway. For most functional pathways PlasmoINT showed a higher number of linkages in both precision networks compared to PlasmoMAP. This increased tendency is more pronounced in the 90% confidence network which furter suggest the improved performance of PlasmoINT in comparison to PlasmoMAP. If only considering the shared genes in each pathway of both networks, most pathways have less number of linkages (data not shown). This suggests that PlasmoINT has eliminated a considerable number of false positive linkages that persist in PlasmoMAP. Panel **C.** depicts the coverage of proteins in the KEGG, GO and MPMP functional pathways by both the PlasmoMAP and PLasmoINT high confidence (90%) network. PlasmoINT contained a higher number of genes for essentially all pathways.

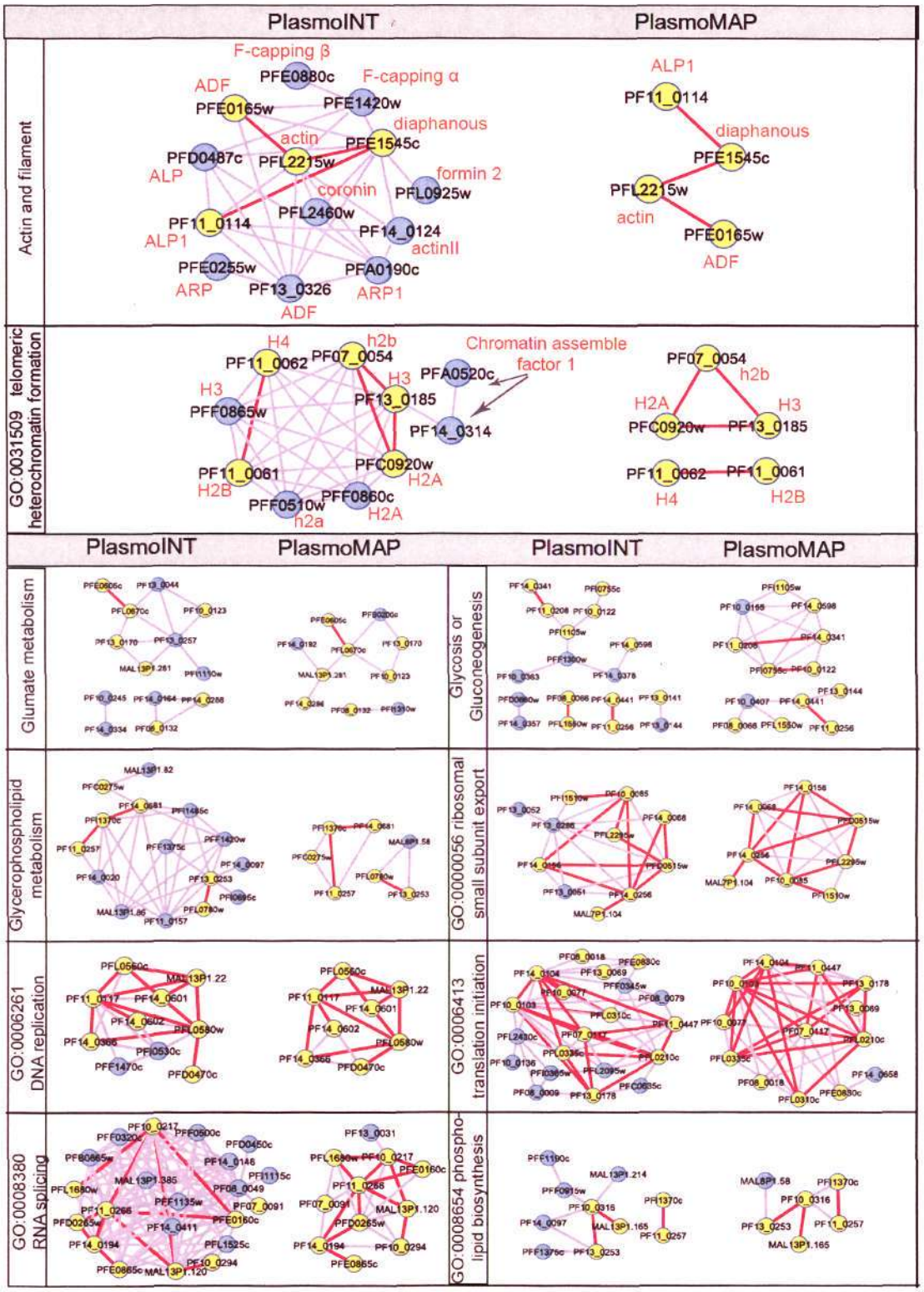


Figure S3.6 Examples of the functional pathway subnetworks from 90% confidence networks of PlasmoINT and PlasmoMAP. 10 metabolic or cellular pathways were reconstructed as subnetworks from the 90% confidence network of both PlasmoINT and PlasmoMAP. The yellow circles and the red edges represent proteins and linkages present in both networks and the purple circles and edges

represent genes and linkages found in each particular network only. For every pathway, PlasmoINT provides substantially higher coverage for both genes and gene linkages. In particular, for the functional group associated with actin and filament formation, (GO:0031509, top panel), PlasmoINT covers 13 genes (in contrast to 4 genes in PlasmoMAP) of that include well established proteins(Baum, et al., 2006): actin complex proteins (actin, actin-II, ARP and ARP1), actin related protein(ALP and ALP1), F-capping complex (F-capping α and β), ADF(PFE0165w and PF13_0326), formin proteins(formin2 and diaphanous) and coronin protein(PFL2460w). In the functional group of "telomeric heterochromatin formation", (GO:0031509, second top panel) PlamsoINT covers 10 genes that include all 8 components of histone complex and two chromation assemble factors that are predicted to represent the core chromatin assembly factors. The 90% confidence PlasmoMAP uncovered only five histone genes.

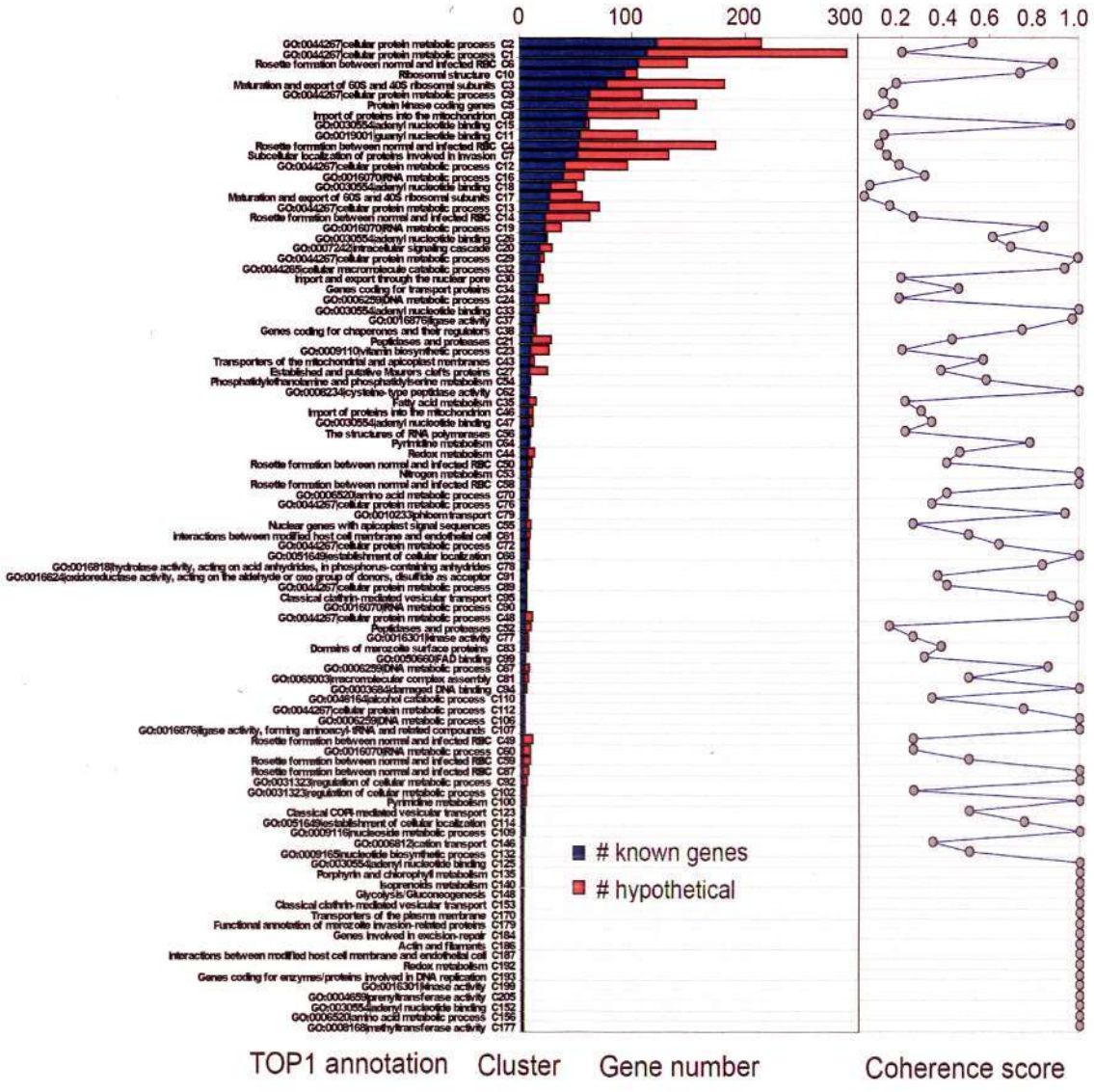


Figure S3.7 Histogram of 105 modules with functional predictions generated from the 90% confidence network. Total 208 modules were identified by MCL (Markov Chain cLuster) method, among which 105 modules (Cluster ID) were contained functionally annotated genes. The modules were ordered according the number of characterized genes (blue bar) and plotted with a number of functionally uncharacterized genes (left panel, orange bar) and the coherence score (right panel). The coherence score were calculated as the fraction of gene pairs that share functional annotations in a given module. Overall, 35 modules contain more than 5 annotated genes and exhibit coherence score > 0.4 (figure 3A). In addition, 58 modules have more than 2 genes and higher coherence > 0.4 (data not shown).

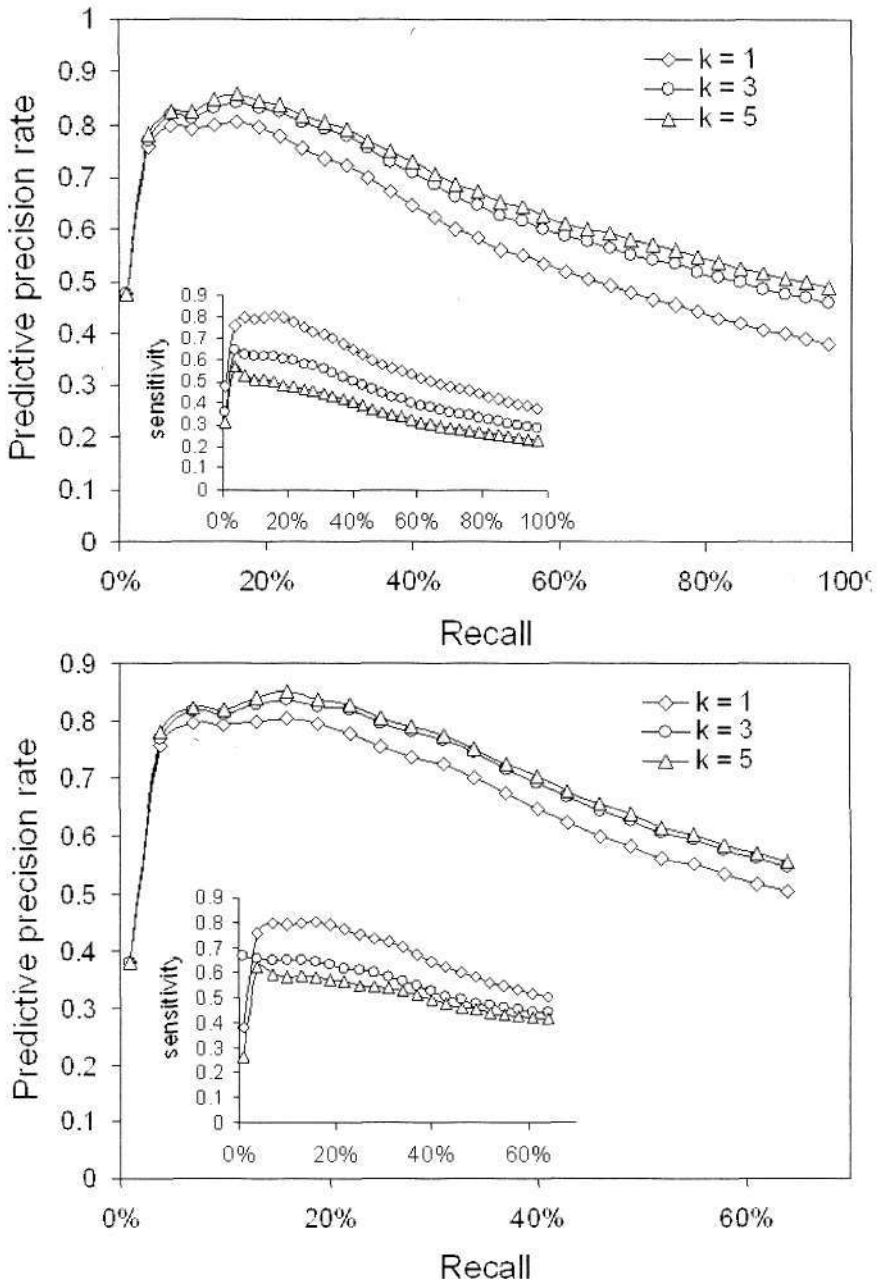


Figure S3.8. The precision rates of network-based predictions of gene function in *P. falciparum* using "leave-one-out" test using the threshold of prediction score. The precision rates of gene functional predictions were plotted against the recall percentage for different k values. K is the number of top assignments of network-based predictions for each gene. When the threshold of prediction score was set at 0.14, TOP 1 assignment has 50% predictive precision rate.

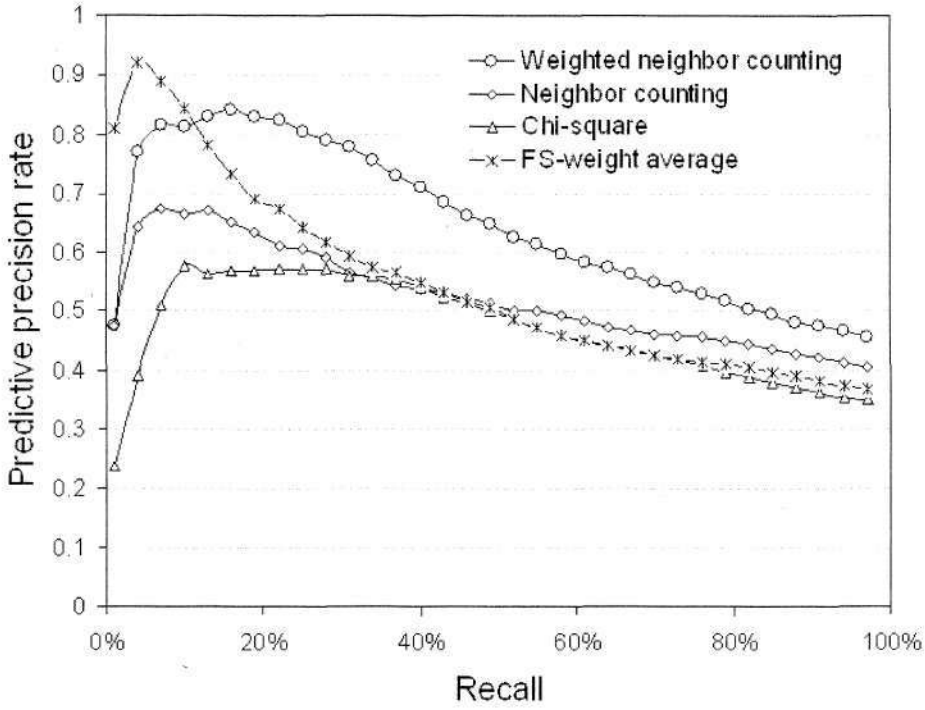


Figure S3.9. Comparisons of the prediction precision rates of different computational methods using "leave-one-out" test. Based on top 3 assignments, the weighted neighbor counting method has significantly higher overall prediction precision rates than those generated by the (simple) Neighbor Counting, Chi-square (Hishigaki, et al., 2001) and FS-weight average (Chua, et al., 2006) methods.

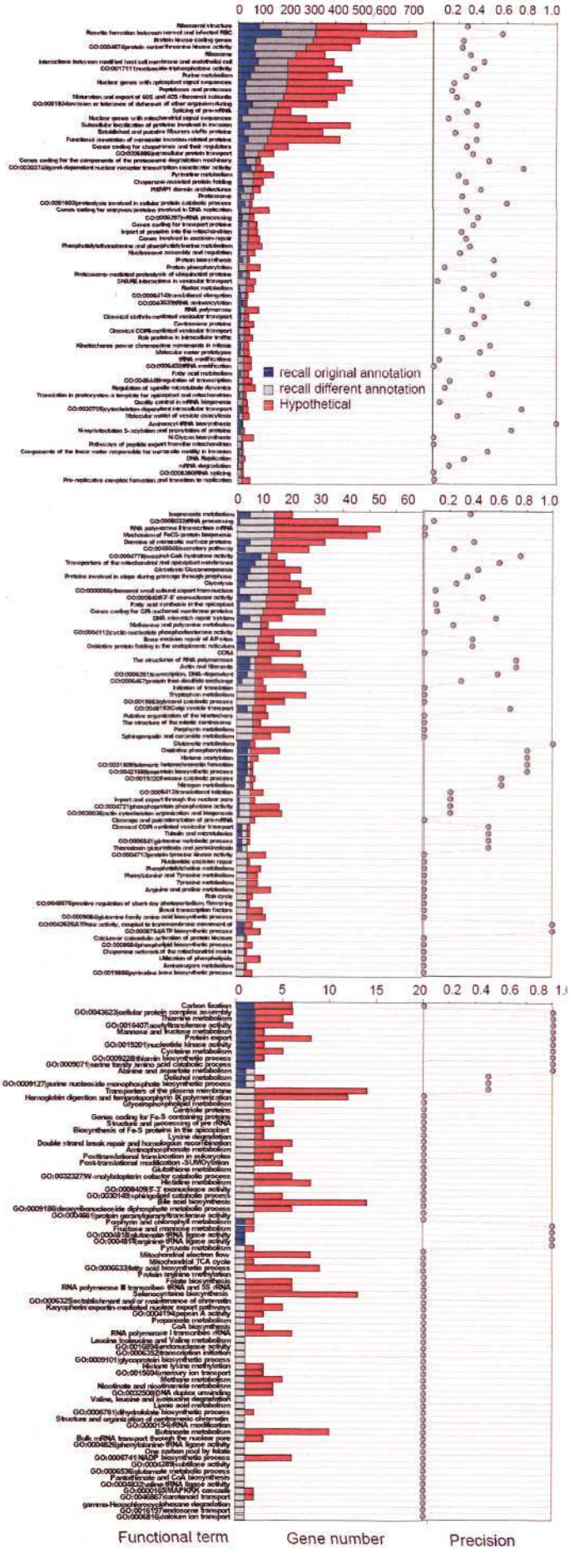


Figure S3.10. Summary of the gene functional prediction precision by WNC in different functional pathways from GO, KEGG, MPM (continuation of figure 3.3B). According to the Top 3 predictions by WNC, the number of genes recalling their original annotations (blue bar), recalling different annotations (gray bar) and hypothetical genes (orange bar) were grouped into functional gene groups (according the newly recalled annotations). The gene counts (left panel) were plotted together with the corresponding prediction precision rates (right panel) for each functional group/pathway. The top 3 prediction methods causes slight underestimation of the precision rates because the "wrong" recall in the pathway classifier can often have other two prediction terms which are correct (see figure S3.8 and S3.9).

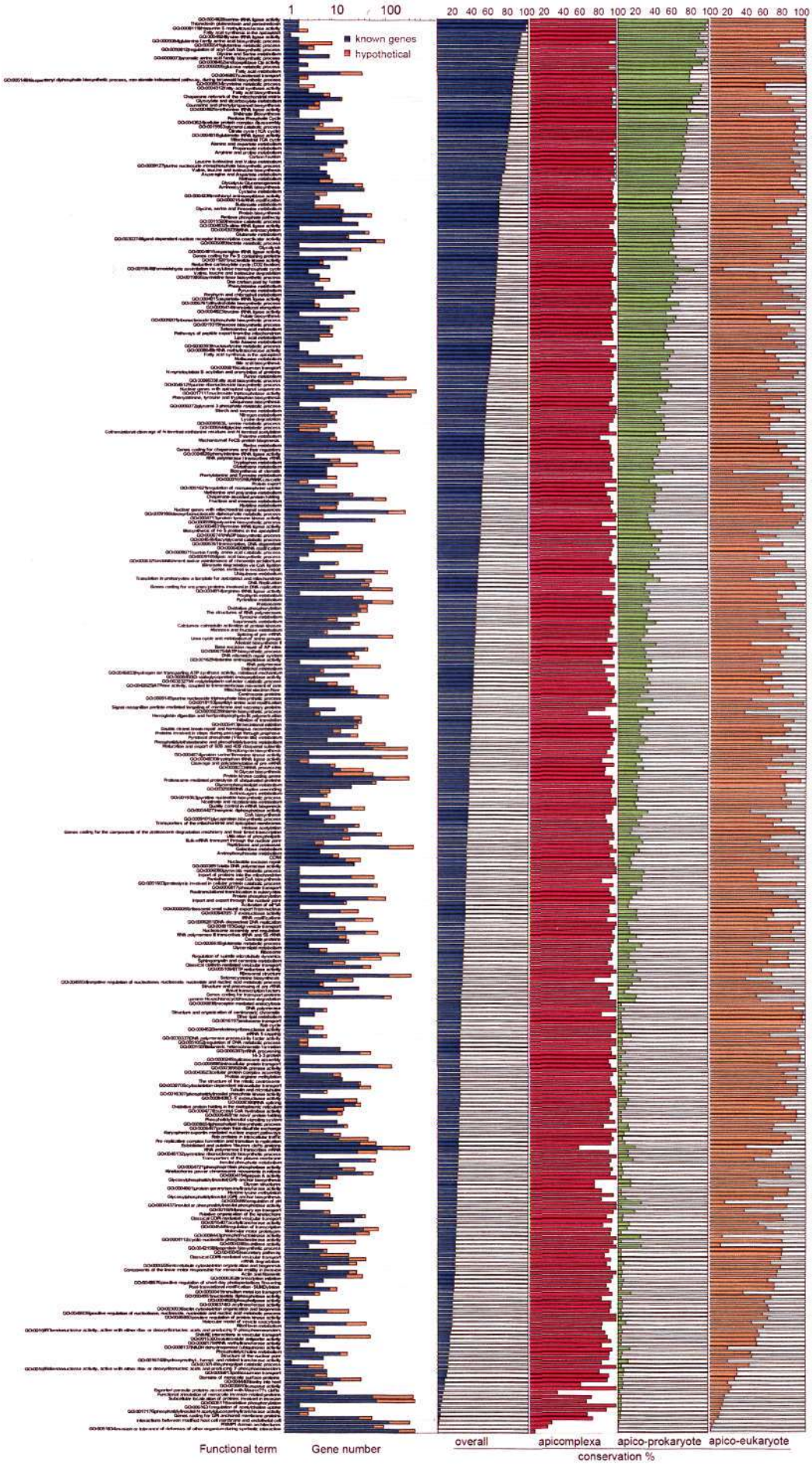


Figure S3.11. Conservation of *P. falciparum* functional pathways among prokaryotes and eukaryotes (continuation of figure 3.3C). The total number of genes annotated by KEEG, GO and MPM (blue bar) and hypothetical genes predicted by WNC (orange bar) were plotted for all 330 functional pathways predicted by the three databases. The functional gene groups were ordered according the overall level of conservation that is calculated as the fraction of the number homologues (reciprocal BLASTP hit, e-value greater than 1e-10) among the total 210 prokaryotic and eukaryotic species (the second panel, blue bar). We also calculated the conservation of different functional pathways in apicomplexa only (third panel, red bar), prokaryotes and apicomplexa (forth panel, green bar) and eukaryote and apicomplexa (right panel, orange bar). The vast majority of P. falciparum pathways are well conserved in other apicomplexans and other eukaryotic species, and a substantial fraction of these are also conserved amongst prokaryotes. The latter class of pathways typically represents basic metabolic functions such as glycolysis, Redox metabolism, fatty acid synthesis or TCA cycle.

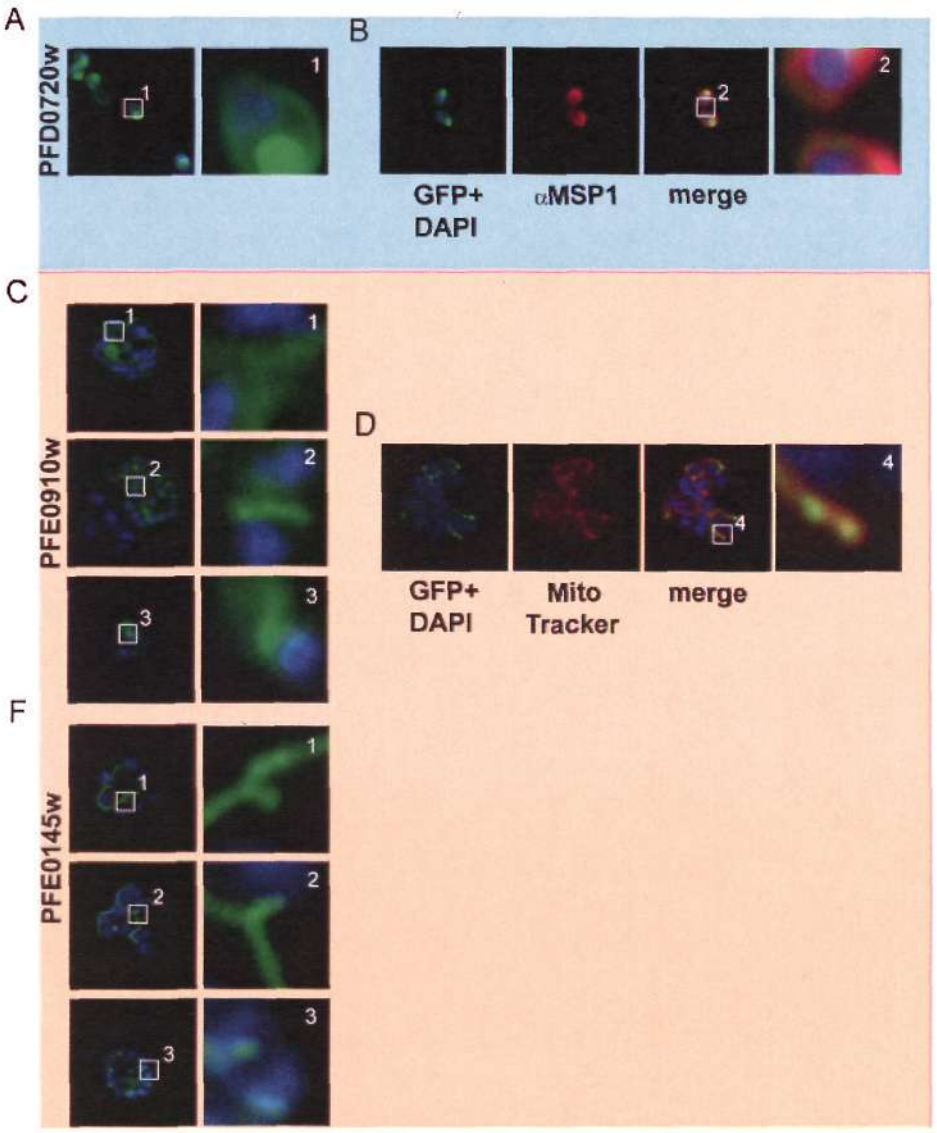


Figure S3.12 Subcellular distribution of the apical protein PFD0720w and the apicoplast and mitochondrion associated proteins PFE0910w. (A-B) In addition to its apical pool, PFD0720w (green) showed in free unfixed (A) and fixed (B) merozoites a faint peripheral distribution that appears to be distinct from the surface marker MSP-1 (red). The boxed regions are depicted in higher magnification and labelled with numbers, nuclei are stained with DAPI (blue). (C-D) While PFE0910w revealed a localisation and dynamic (C1-3) previously described for mitochondrial proteins (van Dooren, et al., 2005) in live parasites and colocalized with the mitochondrium dye MitoTracker, PFE0145w revealed a subcellular distribution and dynamic (D1-3) that is known for apicoplast protein (Waller, et al., 1998).

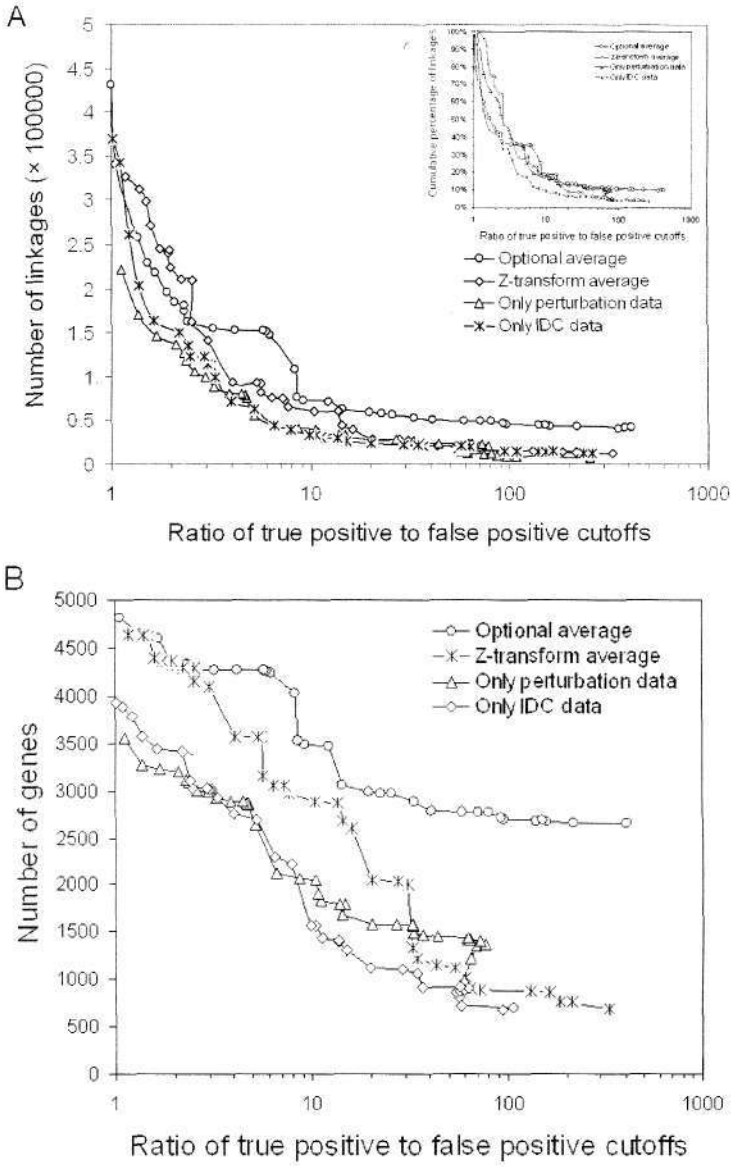


Figure S13. Comparison of four 50% precision rate networks reconstructed different microarray input data. The interactome networks were constructed using different set of the microarray input data (other input dataset remained unchanged): (i) the IDC data only, (ii) the perturbation data only, (iii) the IDC and perturbation data integrated by a z-transform method and (iv) the IDC and perturbation data integrated by the optional average method derived in during work. A. The total number linkages as well as their cumulative percentage (inset) was consistently higher for the optional average method at any given threshold of true-to-false-positive ratios. B. Similarly the optional average method yielded higher gene coverage in different cutoffs of true-to-false-positive ratios.

Table S3.1 Assessment of the accuracies of different data sets in the Bayesian scoring framework based the KEGG benchmark.

Pearson correlation	Benchmark overlap		TP/FP	LS
	positive	negative		
0.9	173	11	26.667	154.18
0.8	933	416	3.495	17.194
0.7	989	1507	1.196	4.234
0.6	816	2537	0.705	2.340
0.5	751	3171	0.489	1.331
0.4	657	3854	0.390	1.079
0.3	530	4343	0.326	0.846
0.2	582	4721	0.282	0.673
0.1	468	4929	0.252	0.568
0	567	5943	0.194	0.729
-0.1	402	4561	0.187	0.541
-0.2	415	3895	0.182	0.600
-0.3	395	3453	0.178	0.512
-0.4	334	2916	0.175	0.645
-0.5	211	2058	0.174	0.646
-0.6	161	1363	0.173	0.623
-0.7	63	602	0.173	1.163
-0.8	11	58	0.173	1.927
-0.9	0	0	0.173	0
Sum	8444	50234		

Mutual information	Benchmark overlap		TP/FP	LS
	positive	negative		
1.4	2	0	-	-
1.3	6	3	2.667	12.130
1.2	31	17	2.045	11.811
1.1	110	66	1.891	11.209
1	270	376	0.993	4.867
0.9	540	1319	0.557	2.453
0.8	873	3091	0.383	1.721
0.7	1226	4861	0.320	1.533
0.6	990	5508	0.272	1.111
0.5	799	6339	0.232	0.799
0.4	918	7345	0.203	0.726
0.3	1058	9211	0.182	0.674
0.2	950	8862	0.169	0.656
0.1	495	4246	0.165	0.743
0	78	634	0.165	0.759
sum	8346	60224		

	Benchmark overlap		TP/FP	LS
	positive	negative		
Protein-protein interaction	3	8	0.333	1.863

Domain interaction (evidence score of Lee et al.)		Benchmark overlap		TP/FP	LS
	log10	positive	negative		
>=100	>=2	400	3	150.33	113.04
31.6	1.5	922	12	92.25	43.162
10	1	136	13	33.8	4.733
3.16	0.5	546	33	13.088	0.787
1	0	425	387	4.386	0.839
<= 0.316	<= -0.5	1106	996	1.330	0.522
sum		3535	1444		

Table S3.2 The 25 core proteins involved in invasion process.

Gene ID	Gene name
PFD0295c	apical sushi protein, ASP
PF11_0344	apical membrane antigen 1, AMA1
MAL7P1.229	Cytoadherence linked asexual protein
PFB0935w	cytoadherence linked asexual protein 2
PFC0110w	Cytoadherence linked asexual protein 3.1
PFI1730w	cytoadherence linked asexual protein 9(CLAG9)
PFC0120w	Cytoadherence linked asexual protein, 3.2
MAL7P1.176	erythrocyte binding antigen
MAL13P1.60	erythrocyte binding antigen 140
PFD1155w	erythrocyte binding antigen-165
PFA0125c	erythrocyte binding antigen-181
PFI1445w	High molecular weight rhoptry protein-2
PF10_0281	hypothetical protein
PFD0110w	normocyte-binding protein 1, pseudogene
MAL13P1.176	*Plasmodium falciparum* reticulocyte binding protein 2, homolog b
PFL2520w	*Plasmodium falciparum*, reticulocyte binding-like protein, homolog 3
PFL0870w	Plasmodium thrombospondin-related apical membrane protein, PTRAMP
PF13_0198	reticulocyte binding protein 2 homolog a
PFD1150c	reticulocyte binding protein homolog 4, Rh4
PFD1145c	reticulocyte binding protein homolog 5, Rh5
PFI0265c	RhopH3
MAL7P1.208	rhoptry-associated membrane antigen, RAMA
PF14_0102	rhoptry-associated protein 1, RAP1
PFE0080c	rhoptry-associated protein 2, RAP2
PFE0075c	rhoptry-associated protein 3, RAP3
PF11_0381	subtilisin-like protease 2