# A protein interaction network of the malaria parasite *Plasmodium falciparum*

Douglas J. LaCount[1,2]*, Marissa Vignali[2]*, Rakesh Chettier[3], Amit Phansalkar[3], Russell Bell[3], Jay R. Hesselberth[2], Lori W. Schoenfeld[1,2], Irene Ota[3], Sudhir Sahasrabudhe[3], Cornelia Kurschner[3], Stanley Fields[1,2] & Robert E. Hughes[3]†

***Plasmodium falciparum* causes the most severe form of malaria and kills up to 2.7 million people annually[1]. Despite the global importance of *P. falciparum*, the vast majority of its proteins have not been characterized experimentally. Here we identify *P. falciparum* protein–protein interactions using a high-throughput version of the yeast two-hybrid assay that circumvents the difficulties in expressing *P. falciparum* proteins in *Saccharomyces cerevisiae*. From more than 32,000 yeast two-hybrid screens with *P. falciparum* protein fragments, we identified 2,846 unique interactions, most of which include at least one previously uncharacterized protein. Informatic analyses of network connectivity, coexpression of the genes encoding interacting fragments, and enrichment of specific protein domains or Gene Ontology annotations[2] were used to identify groups of interacting proteins, including one implicated in chromatin modification, transcription, messenger RNA stability and ubiquitination, and another implicated in the invasion of host cells. These data constitute the first extensive description of the protein interaction network for this important human pathogen.***

The 80% AT content of the *P. falciparum* genome[3] hinders protein expression in heterologous systems[4] and limits both conventional biochemical approaches and comprehensive analyses of this organism's proteins. We overcame this problem by applying a yeast two-hybrid approach that makes use of protein fusions carrying three elements: the Gal4 DNA-binding or activation domain, a random fragment of a *P. falciparum* protein, and an enzyme that allows the growth of auxotrophic yeast deleted for the cognate gene (Supplementary Fig. 1). This procedure was designed to select only those yeast transformants that contained plasmids encoding in-frame and expressed fragments of *P. falciparum* proteins. In addition, this procedure enabled us to generate libraries of DNA-binding domain fusions ('baits') from which randomly chosen transformants could be individually screened in parallel against a library of activation domain fusions ('preys'). Our libraries were derived from RNA isolated from mixed intra-erythrocytic-stage parasites—the stage responsible for pathogenesis in humans—and thus lack genes expressed exclusively in the liver, gametocyte or mosquito stages. Sequence analysis of the inserts from 4,456 DNA-binding domain fusions identified more than 2,000 non-overlapping gene fragments representing 1,295 different *P. falciparum* genes expressed throughout the intraerythrocytic cycle (Supplementary Fig. 2a), indicating that our libraries are complex. Because relatively small fragments of *P. falciparum* genes (about 450 base pairs on average) were cloned, narrowly defined protein–protein interaction domains could be identified.

Coverage of the proteome was obtained by performing more than 32,000 yeast two-hybrid screens, of which 11% yielded positives in which the identities of both interacting protein fragments were determined (Supplementary Table 1). Because the complete set of putative interactions contains both true and false positives, we first sought to eliminate the most obvious class of false positives, namely those protein fragments with many partners, which seem to be 'promiscuous' in the two-hybrid assay. Indeed, although the vast majority of fragments identified relatively few partners, some interacted with many partners (up to a maximum of 207). To identify promiscuous protein fragments, we sequentially applied $k$-means clustering analysis (with $k = 2$) to prey and bait fragments to define two populations based on the number of interacting partners. This approach identified 13 promiscuous prey fragments with more than 31 partners, and 28 promiscuous bait fragments with more than 25 partners (Supplementary Table 2), resulting in the removal of 2,155 interactions involving these fragments from the data set. Because the remaining interactions are listed as pairs of interacting proteins rather than fragments, the total number of partners for a given protein can exceed the thresholds used to remove promiscuous fragments if the protein contains multiple non-promiscuous fragments. Although this analysis removes a significant number of non-specific interactions, other classes of false positive, including promiscuous fragments that resulted in fewer interactions than our threshold values and two-hybrid pairs resulting from mutations in the plasmids or reporter strain, remain in the data set. False positives have been noted in other high-throughput two-hybrid data sets (for example those for proteins of *Drosophila*[5,6]), and can make up a substantial portion of the reported interactions.

The 2,846 unique pairwise interactions that remain constitute a core data set (Supplementary Table 3, also available from the PlasmoDB (http://www.plasmodb.org/) and BIND (http://bind.ca/) databases) that includes 25% of the predicted *P. falciparum* proteins and forms a highly interconnected, scale-free network[7] containing 1,267 proteins linked by 2,823 interactions. An additional 41 proteins are present in small groups of one or two interactions. All categories of proteins seem to have been sampled approximately in proportion to their representation in the genome (Supplementary Fig. 2b). The core data set includes 23 interactions that were previously observed either in *Plasmodium* or between orthologous proteins (Supplementary Table 4). In all, 82% of the interactions include at least one protein annotated as 'hypothetical,' and 33% of the interactions include two hypothetical proteins. The difficulties in expressing *P. falciparum* proteins in heterologous systems precluded experimental confirmation of the interactions by another method. To circumvent this problem, we used several independent bioinformatic analyses to uncover biologically interesting regions of the network.

[1]Howard Hughes Medical Institute, [2]Departments of Genome Sciences and Medicine, University of Washington, Box 357730, Seattle, Washington 98195, USA. [3]Prolexys Pharmaceuticals, Inc., 2150 West Dauntless Avenue, Salt Lake City, Utah 84111, USA. †Present address: Buck Institute, Novato, California 94945, USA.
*These authors contributed equally to this work.

Because high degrees of local network interconnectivity can identify sets of functionally related proteins[8], we surveyed the network for groups of proteins with a greater number of connections than would be expected by chance. We parsed the network into 1,308 primary subnetworks containing a protein, its direct binding partners and all interactions between them, and calculated a connectivity coefficient (defined as the number of interactions divided by the number of proteins). A comparison of the distribution of connectivity coefficients present among the experimentally observed subnetworks with those derived from randomized subnetworks of equal size showed an enrichment for highly connected subnetworks in the real data (Fig. 1a); 96 subnetworks showed a higher degree of interconnectivity than expected by chance ($P \leq 0.05$) when compared with the mean of the connectivity coefficients from 100 randomized subnetworks of the same size (Supplementary Table 5). Several of these 96 subnetworks shared a common set of interactions that revealed a region of the interaction network with a high degree of interconnectivity. This region overlaps the complex with the highest connectivity in the data set as defined by MCODE[9], an algorithm that identifies densely connected areas of protein interaction networks (Supplementary Fig. 3, complex 1). On the basis of these analyses, we identified a group of interacting proteins likely to integrate chromatin modification, transcriptional regulation, mRNA stability and ubiquitination (Fig. 1b). Whereas 11% of the interactions in the whole data set were observed in two or more independent experiments, more than 40% of the interactions in this group were observed in multiple independent experiments.

This set of interacting proteins is centred on PF08_0034, the *P. falciparum* orthologue of the yeast histone acetyltransferase Gcn5. The interactions established by PfGcn5 are mediated by an amino-terminal extension that is absent from the yeast Gcn5 (ref. 10), indicating that this group might represent a *Plasmodium*-specific pathway that regulates gene expression. Other potential chromatin-modifying proteins in this group include PFF1440w, a protein containing a PHD domain, bromodomain and SET domain that potentially recognizes acetylated nucleosomes and acts as a histone methyltransferase[11,12], and PFF1470c, an orthologue[13] of yeast DNA Pol2 (Polε), which is involved in DNA replication and chromatin silencing at telomeres[14]. Two putative transcription factors are also present in this group: PF11_0241, which contains a Myb DNA-binding domain (and directly interacts with both PF08_0034 and PFF1440w), and PF10_0075, which has a DNA-binding AT-hook and a BTB/POZ domain, which is found in some transcription factors involved in the recruitment of histone deacetylase complexes[15]. These interactions indicate that chromatin-modifying complexes might be targeted to specific regions in the genome to regulate transcription and are of particular significance given the apparently unique features of gene expression of the parasite[16,17] and the dearth of recognizable transcription factors encoded by its genome[18,19]. The presence of three ubiquitin metabolism proteins (the RING finger and forkhead-associated (FHA)-domain protein PFL0275w, the HECT-domain protein PFF1365c and the UCH-domain protein PFI0225w) indicates that ubiquitination might be involved in regulating the stability or activity of proteins in this group; Gcn5-containing HAT complexes can deubiquitinate histones[20]. This group also contains MAL8P1.104, the *P. falciparum* orthologue of Caf1, the major mRNA deadenylase in yeast and a member of the Ccr4–Not complex. Indeed, this group seems analogous to Ccr4–Not, which also integrates transcription regulation, chromatin modification, ubiquitination and RNA stability[21]. Other interactions implicated in *Plasmodium* nucleic acid metabolism[18,19] are shown in Supplementary Fig. 4.

Interacting proteins from *S. cerevisiae* tend to have similar mRNA abundance profiles[22]; positively correlated mRNA expression is therefore generally taken as an indication that a putative interaction is more likely to be real. However, this correlation is not as evident in the large-scale *C. elegans* protein interaction data[23], and two genes need not share similar expression profiles for their proteins to be present in the cell at the same time. For example, one gene may be expressed constitutively whereas the other is induced under certain conditions, or their proteins may have different half-lives. Indeed, the time of maximal accumulation for a substantial portion of *P. falciparum* proteins is shifted relative to the time of maximum mRNA abundance[24]. To examine the relationship between *P. falciparum* protein interaction data and mRNA abundance, we compared our core data set to data from two genome-scale gene expression studies that addressed the timing of mRNA accumulation during the *P. falciparum* life cycle[16,17]. We calculated Pearson correlation coefficients (PCCs) for each protein pair and averaged these values for proteins with more than one partner. We identified 82 proteins with average PCCs significantly higher than expected compared with mean PCCs from 100 randomizations in which the identity of the partners was changed but the total number of partners remained constant ($P \leq 0.05$; Supplementary Table 6). Several of
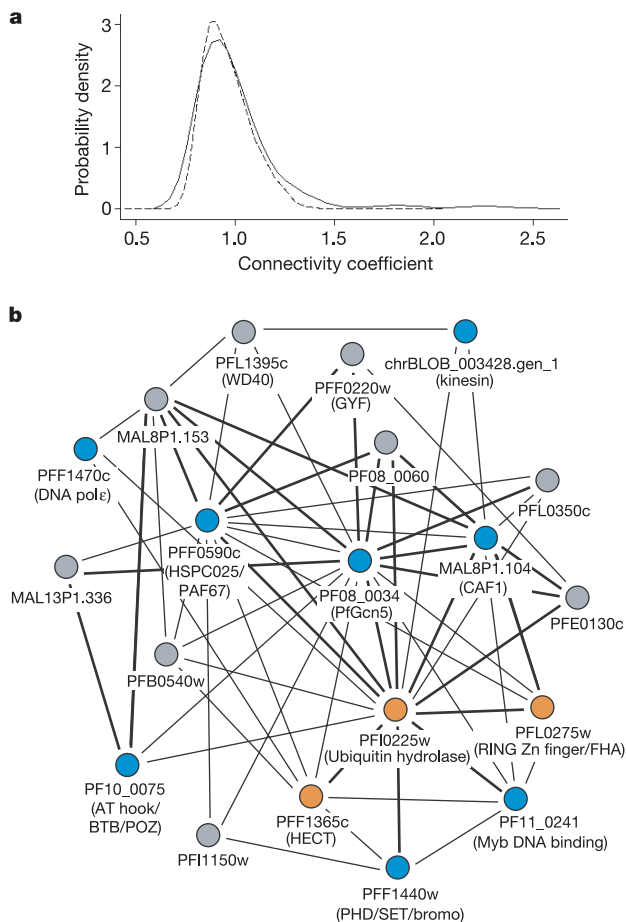
**Figure 1 | Connectivity analysis. a,** Distribution of connectivity coefficients from experimental (solid lines) and randomized (dashed line) subnetworks containing five or more proteins (a central node plus four or more partners). **b,** The most highly connected region of the *P. falciparum* protein interaction network. The graph shows interactions (lines) between proteins (circles) involved in transcription or chromatin metabolism (blue) and ubiquitin metabolism (orange); proteins with no additional supporting evidence linking them to these processes are shown in grey. Thin lines indicate interactions observed in a single yeast two-hybrid experiment; thick lines show interactions found in reciprocal orientation or in two or more independent yeast two-hybrid screens. Systematic gene names are as given in PlasmoDB; when available, common names, putative functions and domains are shown in parentheses. Only proteins with two or more interactions with other group members are shown; additional partners for these proteins are listed in Supplementary Table 3.

these 82 proteins are expressed during schizogony, the time at which new merozoites are being formed. In addition, when we compared the interaction data set with the gene clusters defined in ref. 17 based on mRNA abundance profiles, we identified cluster 15 as having a higher than expected number of interactions among proteins within the cluster (19 interactions observed, 10 predicted from a randomized network, $P < 0.05$; Supplementary Table 7).

Cluster 15 contains proteins implicated in the invasion of host cells, including merozoite surface protein 1 (MSP1, PFI1475w), an essential protein that coats the surface of merozoites and is thought to be required for the invasion of red blood cells. This potential vaccine candidate has several conserved blocks of sequence, some of which establish interactions with uncharacterized, coexpressed proteins that might also have a function in the invasion of host cells (Supplementary Fig. 5). To identify other potential host-cell-invasion proteins, we screened our data set for pairs of interacting proteins with expression patterns similar to that of MSP1 (that is, expression peaking at the time when new merozoites are formed and repressed in the early phase of the 48-h infection cycle). We identified 103 interactions among 89 proteins, 75 of which are linked together in an extended region of the network (Supplementary Fig. 6). The core of this set of proteins (Fig. 2) recapitulates a previously shown interaction between MSP1 and MSP9 (PFL1385c)[25] and links 19 uncharacterized proteins to 16 proteins that are involved in the invasion of host cells or are localized to the merozoite surface. Of the 19 uncharacterized proteins, 6 have been detected in merozoites by mass spectrometry[26] and 4 have predicted signal peptides[27]. Consistent with PFD0230c (a putative type I dipeptidyl aminopeptidase that interacts with MSP3) having a function in the invasion of host cells is the observation by M. Klemba and D. Goldberg (personal

communication) that it seems to localize to the apical region of developing merozoites. Other interactions link merozoite surface proteins (MSP1 and MSP9) to proteins localized to rhoptries (RhopH1, RhopH2 and RhopH3), indicating the potential for transient interactions that occur during the invasion of host cells after the contents of the rhoptries have been released.

Because specific protein domains are often associated with discrete biological processes, enrichment of particular domains in subnetworks can implicate proteins relevant to a process. Similarly, enrichment of proteins sharing common Gene Ontology (GO) annotations (http://www.geneontology.org) can also implicate proteins from a subnetwork in biological processes. We therefore searched the *P. falciparum* interaction network for primary subnetworks
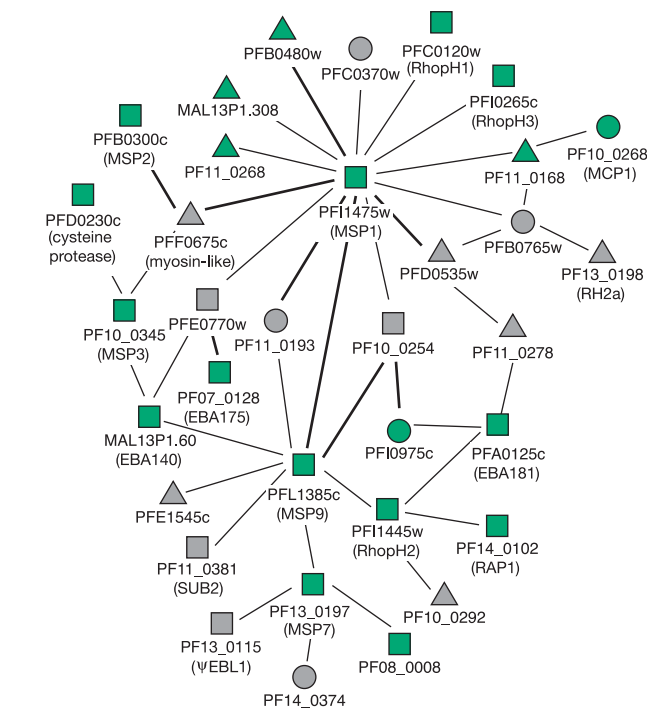


**Figure 2 | Interactions between uncharacterized *P. falciparum* proteins and proteins involved in the invasion of host cells.** Gene names and thin and thick lines are as in Fig. 1. Squares, proteins with a predicted signal peptide; triangles, at least one predicted transmembrane domain; circles, no predicted transmembrane domains; green, peptides from that protein were detected by mass spectrometry in merozoites[26]; grey, protein not detected in merozoites. Only proteins with mRNA expression profiles similar to that of MSP1 are shown; additional partners for these proteins are listed in Supplementary Table 3.
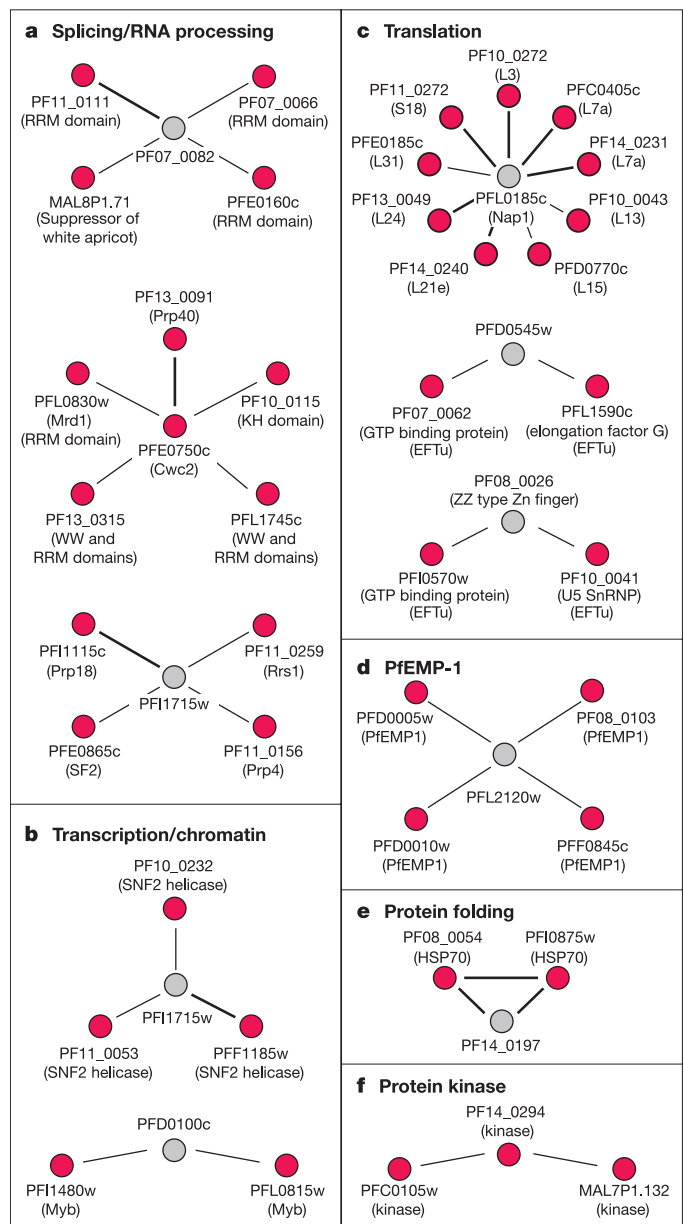


**Figure 3 | Subnetworks with shared protein domains and GO annotations.** Gene names and thin and thick lines are as in Fig. 1; when available, common names, putative functions or protein domains are shown in parentheses. Nodes coloured red share the feature indicated at the top of the panel; nodes coloured grey do not. Only partners of the central protein that contain the indicated feature are shown; additional partners for these proteins are listed in Supplementary Table 3.

associated with protein domains or GO annotations (Supplementary Tables 8 and 9). We identified several subnetworks with an enrichment of RNA recognition motifs (RRM; Fig. 3a), indicating that the proteins in these subnetworks might be involved in RNA processing or splicing. For example, PF07_0082, a hypothetical protein, interacts with three proteins that have RRM domains and a fourth that might be an orthologue of the *Drosophila* splicing factor suppressor of white apricot. PFE0750c, a potential orthologue of the *S. cerevisiae* splicing factor Cwc2, interacts with four putative RNA-binding proteins and an orthologue of the yeast splicing protein Prp40. PFI1715w binds to four putative RNA-processing proteins, including orthologues of yeast splicing factors Prp4 and Prp18, and human alternative splicing factor 2; the putative *P. falciparum* Prp4 protein in turn binds to the *P. falciparum* orthologue of yeast Prp9 (PFI1215w, not shown). PFI1715w also interacts with three proteins bearing SNF2 helicase domains (Fig. 3b). Given that proteins containing SNF2 domains are involved in chromatin remodelling, PFI1715w might provide a link between gene expression and splicing in *P. falciparum*. In some cases, domain and GO annotation enrichment indicate unexpected alternative functions for *P. falciparum* proteins. We found an orthologue of the yeast nucleosome assembly protein Nap1 (PFL0185c) in interactions with nine ribosomal proteins (Fig. 3c) and a protein involved in ribosome biogenesis (PF11_0259, Rrs1; not shown), possibly indicating a role for PFL0185c in ribosome assembly or translation. Similarly, a putative subunit of the N-terminal acetyltransferase complex (PFL2120w) interacts with the cytoplasmic tails of four PfEMP-1 (*P. falciparum* erythrocyte membrane protein-1) proteins (Fig. 3d) and two other proteins that are known or predicted to be exported to the host cell cytoplasm (MAL7P1.170 and PFA0110w, not shown), indicating that acetylation (or binding to PFL2120w) might be involved in protein trafficking to this compartment. Other examples of domain enrichment implicate uncharacterized *P. falciparum* proteins in protein folding (Fig. 3e) and a possible protein kinase signalling cascade (Fig. 3f).

Last, we compared our set of interactions with the more than 150 parasite-derived proteins known or predicted to be exported into the host cell cytoplasm[28–30]. These proteins extensively modify infected red blood cells, establishing a new secretory system in the cytoplasm and generating knob-like structures on the cell surface. We identified 15 interactions among 19 exported proteins, which might provide insight into the structure of parasite-mediated modifications of the host cell (Supplementary Fig. 7).

The data set of putative protein interactions described here greatly exceeds the number of previously known interactions for *P. falciparum* and provides the basis for focused experiments on a variety of biological processes. As this network reflects pathways and processes of the parasite, this information should be relevant both to our understanding of the basic biology of the organism and to the discovery of new drug and vaccine targets. Although the difficulties in working with this organism currently preclude the types of experimental validation that are available in model organisms, we expect that the accumulation and integration of large-scale gene and protein expression studies and protein interaction data sets will continue to provide informatics-based approaches towards understanding this human parasite.

## METHODS

**Bait and prey construction.** Complementary DNA was generated from poly(A)$^+$ RNA isolated from mixed-staged (strain 3D7)-infected erythrocytes (a gift from K. Ganesan and P. Rathod) and inserted between the Gal4 transcriptional activation domain and the *Schizosaccharomyces pombe URA4* coding region of pOAD.102 (prey plasmid) or the Gal4 DNA-binding domain and the *S. cerevisiae MET2* coding region of pOBD.111 (bait plasmid). Yeast transformed with bait or prey plasmids were plated on medium lacking uracil (prey) or methionine (bait) to select for transformants expressing the markers fused to the cDNA inserts. Additional information about the plasmids, yeast strains and library construction can be found in Supplementary Methods.

**Yeast two-hybrid process.** Individual bait colonies were picked at random and clonally expanded in liquid medium in 96-well plates. Aliquots from the prey libraries were added to each well; mating occurred overnight. Matings were plated on medium that selected for the mating event, the expression of the auxotrophic markers fused to the cDNA inserts, and the activity of the metabolic reporter genes *ADE2* and *HIS3*. The cDNA inserts from yeast that grew on this selection medium were amplified by polymerase chain reaction and then sequenced. The identities of inserts were determined by querying the sequences against the annotated *P. falciparum* genes in PlasmoDB version 4.0 and the genome sequences from PlasmoDB version 3.3 that were excluded from version 4.0. Additional details are provided in Supplementary Methods.

**Removal of false-positive bait and prey fragments.** Activation and DNA-binding domain inserts were treated as protein fragments and independently grouped into two populations by *k*-means clustering on the basis of their number of partners. Interactions involving fragments from groups with the greater number of partners were deemed promiscuous and removed from the final data set. Additional information is provided in Supplementary Methods.

**Computational analysis.** The identification of local regions of the protein interaction network with enhanced connectivity, comparisons of protein interaction data with *P. falciparum* microarray data sets from refs 16 and 17, and the discovery of proteins whose partners were enriched for protein domain or GO annotations were performed as described in Supplementary Methods.

1. Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64**, 1–11 (2001).
2. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
3. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
4. Sibley, C. H. *et al.* Yeast as a model system to study drugs effective against apicomplexan proteins. *Methods* **13**, 190–207 (1997).
5. Formstecher, E. *et al.* Protein interaction mapping: a *Drosophila* case study. *Genome Res.* **15**, 376–384 (2005).
6. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
7. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
8. Rives, A. W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA* **100**, 1128–1133 (2003).
9. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
10. Fan, Q., An, L. & Cui, L. *Plasmodium falciparum* histone acetyltransferase, a yeast GCN5 homologue involved in chromatin remodeling. *Eukaryot. Cell* **3**, 264–276 (2004).
11. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
12. Ragvin, A. *et al.* Nucleosome binding by the bromodomain and PHD finger of the transcriptional cofactor p300. *J. Mol. Biol.* **337**, 773–788 (2004).
13. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
14. Iida, T. & Araki, H. Noncompetitive counteractions of DNA polymerase epsilon and ISW2/yCHRAC for epigenetic inheritance of telomere position effect in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **24**, 217–227 (2004).
15. Huynh, K. D. & Bardwell, V. J. The BCL-6 POZ domain and other POZ domains interact with the co-repressors N-CoR and SMRT. *Oncogene* **17**, 2473–2484 (1998).
16. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1**, E5 (2003).
17. Le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503–1508 (2003).
18. Coulson, R. M., Hall, N. & Ouzounis, C. A. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Res.* **14**, 1548–1554 (2004).
19. Aravind, L., Iyer, L. M., Wellems, T. E. & Miller, L. H. *Plasmodium* biology: genomic gleanings. *Cell* **115**, 771–785 (2003).
20. Daniel, J. A. *et al.* Deubiquitination of histone H2B by a yeast acetyltransferase complex regulates transcription. *J. Biol. Chem.* **279**, 1867–1871 (2004).
21. Collart, M. A. Global control of gene expression in yeast by the Ccr4-Not complex. *Gene* **313**, 1–16 (2003).
22. Ge, H., Walhout, A. J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19**, 551–560 (2003).
23. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
24. Le Roch, K. G. *et al.* Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res.* **14**, 2308–2318 (2004).

25. Li, X. *et al.* A co-ligand complex anchors *Plasmodium falciparum* merozoites to the erythrocyte invasion receptor band 3. *J. Biol. Chem.* **279,** 5765–5771 (2004).
26. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
27. Kissinger, J. C. *et al.* The *Plasmodium genome* database. *Nature* **419,** 490–492 (2002).
28. Cooke, B. M., Lingelbach, K., Bannister, L. H. & Tilley, L. Protein trafficking in *Plasmodium falciparum*-infected red blood cells. *Trends Parasitol.* **20,** 581–589 (2004).
29. Hiller, N. L. *et al.* A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306,** 1934–1937 (2004).
30. Marti, M., Good, R. T., Rug, M., Knuepfer, E. & Cowman, A. F. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306,** 1930–1933 (2004).