RESEARCH ARTICLE

Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle

Karine G. Le Roch,[1]* Yingyao Zhou,[2] Peter L. Blair,[3]
Muni Grainger,[4] J. Kathleen Moch,[3] J. David Haynes,[5]
Patricia De la Vega,[3] Anthony A. Holder,[4] Serge Batalov,[2]
Daniel J. Carucci,[3] Elizabeth A. Winzeler[1,2]*

The completion of the genome sequence for *Plasmodium falciparum*, the species responsible for most malaria human deaths, has the potential to reveal hundreds of new drug targets and proteins involved in pathogenesis. However, only $\sim$35% of the genes code for proteins with an identifiable function. The absence of routine genetic tools for studying *Plasmodium* parasites suggests that this number is unlikely to change quickly if conventional serial methods are used to characterize encoded proteins. Here, we use a high-density oligonucleotide array to generate expression profiles of human and mosquito stages of the malaria parasite's life cycle. Genes with highly correlated levels and temporal patterns of expression were often involved in similar functions or cellular processes.

P. falciparum, the protozoan parasite responsible for the most severe form of malaria, causes 1.5 to 2.7 million deaths every year, mostly among children in sub-Saharan Africa (*1*). The development of resistance in the parasite to effective and inexpensive drugs, the lack of a licensed malaria vaccine, and the fundamental complexity inherent in the malaria parasite mean that there is an urgent need to better understand the function of *P. falciparum* genes and their biological role to support the development of new and effective antimalarial strategies. In 1996, an international consortium was established to determine the sequence of the 26 Mb *P. falciparum* genome, resulting in its publication in 2002 (*2–6*). Mass spectrometry analysis has discerned that at least 2391 genes are transcribed and translated into proteins in one or more stages of the parasite life cycle (*7*, *8*), but the cellular roles of the majority of the proteins need to be elucidated.

Although systems for transient and stable transfection are being developed and will provide important tools to dissect the function of parasite genes (*9–12*), they remain time-consuming with an extremely low efficiency. Therefore, additional new high-throughput methods are needed to collect functional information about genes. In the past few years, gene expression profiling has emerged as an important tool to study gene function in genetically intractable organisms (*13*) and has transformed the traditional gene-by-gene analysis (*14–16*). Here, we describe and use a custom-made, high-density oligonucleotide array ($\sim$500,000 features), designed using the *P. falciparum* genome nucleotide sequence, to determine the relative level and temporal pattern of expression of more than 95% of the predicted *P. falciparum* genes as the parasite moves through its life cycle.

Array design and expression level. Nine different stages during development were examined: mosquito salivary gland sporozoites, which infect humans; seven periodic erythrocytic asexual stage time points, spanning from early ring forms through mature schizonts to free merozoites, which are the stages responsible for the pathological manifestations of malaria; and finally, the sexual stage gametocytes, the form by which the parasite is transmitted from humans to mosquitoes. To study the transcriptome of the malaria parasite, we designed a custom oligonucleotide array containing 260,596 25-nucleotide single-stranded probes from predicted coding sequence (including mitochondrion and plastid genome sequences) and 106,630 probes from noncoding sequence together with the appropriate controls (*17*). Because the *P. falciparum* nuclear genome is relatively small ($\sim$23 Mb), this design allowed one of the 367,226 probes to be placed, on average, every 150 bases on both DNA strands. Total RNA was extracted from sporozoites; from six periodic intracellular asexual blood stages grown in culture and synchronized by means of two independent methods, a 5% D-sorbitol treatment (cell cycle S) (*18*) and a temperature cycling incubator (cell cycle T) (*19*); and from merozoites and mature stage IV and V gametocytes. The use of the two synchronization methods was expected to reveal genes under true cell-cycle control by eliminating variation due to either chemical or temperature stress. The RNA was labeled by a strand-specific protocol and hybridized to the array (*16*). Probes on the array mapping to sequences within the predicted gene models (*17*) were used to compute gene expression levels by means of the match-only integral distribution algorithm (MOID) (*20*). For large genes we only used the twenty probes closest to the $3'$ end. Because the cDNA synthesis and in vitro transcription reactions are initiated at the $3'$ end of an mRNA molecule, inclusion of all probe data could lead to decreased expression values for large transcripts, which are more prone to nicking. Each probe within a set has its own hybridization properties, but changes in expression profile were determined at the probe set level with statistical tools; thus, the level of confidence in the expression pattern for a particular gene is high (fig. S1). Despite the fact that two different synchronization methods were used for the erythrocytic cycle, comparability between experiments was very good. Correlation coefficients of logarithm-transformed expression values for equivalent time points in the cell cycles obtained by the two synchronization methods ranged from 0.87 to 0.91 (fig. S2). Correlation coefficients of the logarithm-transformed expression values between the two sporozoite hybridizations when one amplification step was used was 0.92, showing that the amplification step introduced a minimal degree of bias into the sample. Gene expression data for all stages and for all 5159 genes probed on the array (fig. S3) are available in table S1 and at http://plasmoDB.org. Because of the array design, which includes many probes to putative noncoding regions, the data from these experiments can be reanalyzed if gene models change (fig. S1C). Probe data from noncoding regions were not considered in this analysis, but extensive transcription on the opposite strand of the predicted open reading frames has been identified and described (*21*).

Because our chip contains multiple probes per gene, a quantitative estimate of a gene's transcript abundance could be obtained for genes with multiple probes. Using a probability function based on a gene's expression level ($E > 10$) and its probe signal distribution ($\log P < -0.5$) (fig. S4), we found that 4557 genes (88% of the predicted genes)

[1]Department of Cell Biology ICND202, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. [2]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA. [3]Malaria Program, Naval Medical Research Center, Silver Spring, MD 20910–7500, USA. [4]Division of Parasitology, National Institute for Medical Research, London NW7 1AA, UK. [5]Department of Immunology, Walter Reed Army Institute of Research, Silver Spring, MD 20910, USA.

*To whom correspondence should be addressed. E-mail: winzeler@scripps.edu (E.A.W.); leroch@scripps.edu (K.G.L.)

were expressed in at least one stage of the life cycle (Fig. 1), with expression levels varying by five orders of magnitude. The proportion of expression for different functional classes for different stages (Fig. 2) shows the shift in transcriptional energy from protein synthesis (ring and trophozoite stages) to cell surface structures (schizonts and sporozoites) within the life cycle. Among the most highly expressed genes (2%) were many encoding ribosomal proteins, histones, or actin as well as genes for proteins involved in glucose metabolism. The high level for the latter group in asexual blood stages is consistent with prior studies of glycolysis in this stage (*22*). Among the most highly expressed genes in the erythrocytic cycle, we found four genes coding for early transcribed membrane proteins, including three in the top ten (computed by summing expression levels for all stages). These genes were recently described as a new invasion-related gene family (*etramps* or PfSEP) that includes PFB0120w (PfSEP2), PF11_0040 (PfSEB11-1), and PF11_0039 (PfSEB11-2) (*23*, *24*). The group of 885 genes for which mRNA was not detected during the erythrocytic cycle included 36 transcriptionally silent genes that are likely involved in antigenic variation (such as *rifin* genes) and those that are sporozoite and gametocyte stage specific (*7*). A total of 602 genes were not expressed in any stage examined, and a majority (87%) of these code for hypothetical proteins that may be expressed at other stages of the life cycle, including the mosquito gut or the pre-erythrocytic liver stages.

Cell-cycle regulation and constitutive expression of *P. falciparum* genes. By applying the one-way ANOVA statistical test to identify differentially expressed genes from the time course data (*20*), we found that 43% of the expressed genes were cell-cycle regulated ($P \leq 0.05$). Among them, 1489 genes were found to be regulated in both erythrocytic cell cycle samples (S and T), with an additional 746 genes differentially regulated in the sporozoite and the gametocyte stages (table S2). Regulated genes were required to have a minimum fold change of 1.5 across the life cycle.

In contrast, 51% of the expressed genes were designated as constitutively expressed at the described stages of the life cycle. This group contains genes coding for both uncharacterized hypothetical proteins and housekeeping proteins. Many of the hypothetical proteins in this group are likely involved in maintenance of parasite function throughout the life cycle, and those without human orthologs might represent targets for drug development.

Cluster analysis. One of our goals was to demonstrate that genes with similar functions have similar expression profiles. If such an assumption proved true, then expression profiling could allow us to rapidly and empirically assign functions to the thousands of uncharacterized proteins encoded by the genome. To test this, genes were first grouped on the basis of time of expression throughout the *P. falciparum* life cycle; i.e., expression data for the 2235 genes showing statistically significant variation in transcription were grouped by a robust *k*-means program. The robust *k*-means clustering algorithm runs on top of the standard *k*-means clustering algorithm and consolidates results from 1000 independent standard *k*-means clustering runs. This method eliminates arbitrary gene-cluster associations because of the random selection of the initial clustering centers in the standard *k*-mean clustering algorithm (*17*). We chose $k = 15$ as reasonable for the number of conditions under consideration. The resultant clusters are shown in fig. S5A. Data for other *k* values are given at http://carrier.gnf.org/publications/CellCycle. Although most of the genes belonging to these clusters encode proteins with unknown function (48% to 88%, depending on the cluster) (Table 1), enough have defined cellular roles to allow us to conclude

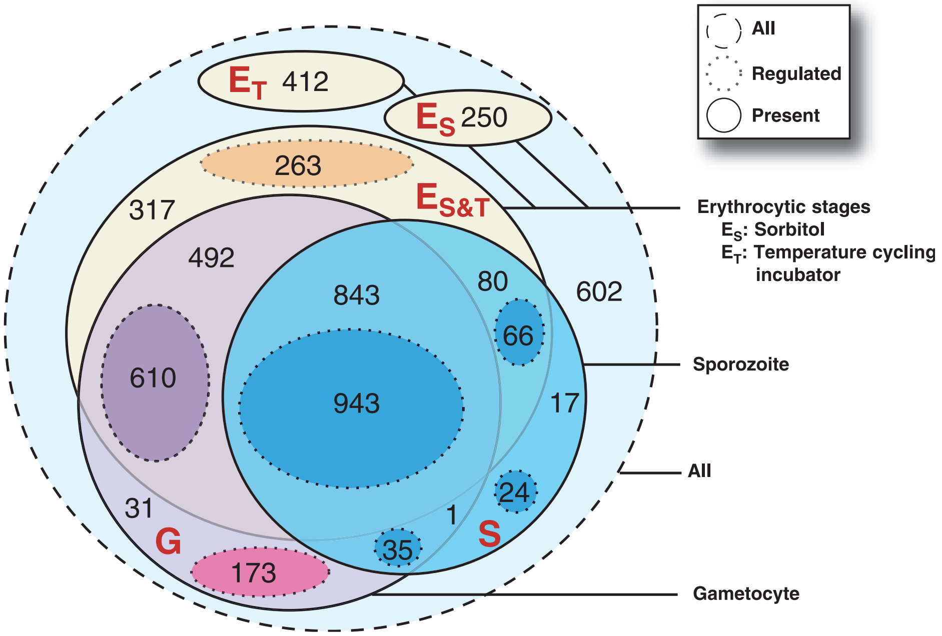

Fig. 1. Venn diagram of life cycle–regulated genes. The solid circles denote genes that are considered expressed in the erythrocytic cell cycle (E_T and E_S), gametocytes (G), or sporozoites (S). The dashed circles represent the subset of present genes that are regulated ($P \leq 0.05$ and fold change $\geq$ 1.5).

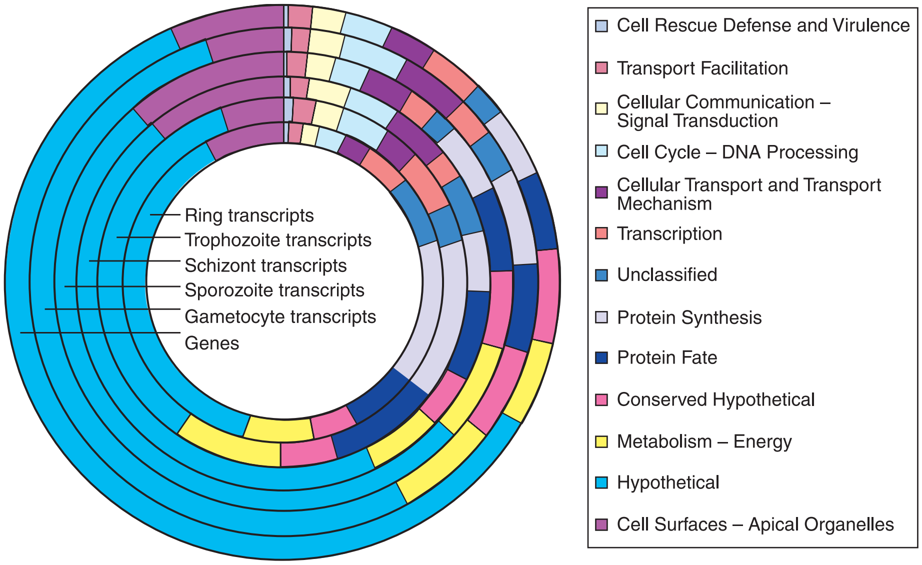

Fig. 2. Distribution of transcripts for several stages. Expression levels for late ring [synchronized by thermocycling (T)], late trophozoite (T), late schizont (T), sporozoite, and gametocyte stages were summed by each MIPs functional category. The outer ring shows the proportion of genes in each category.

that membership in a cluster is highly nonrandom and thus highly predictive. By comparing the published gene ontology rosters with the cluster gene rosters (Table 1), we found that the probability of the degree of overlap we observed having been found by chance was often many orders of magnitude smaller than would be considered acceptable ($P \leq 0.05$). For example, the probability that 13 of the 18 differentially ex-

Table 1. Description of clusters. *P* values were determined from published gene ontologies at http://plasmoDB.org. The number of adjacent pairs indicates the number of genes in a cluster that are physically adjacent to one another (regardless of strand), whereas the number in parentheses indicates the number expected by chance. MAP, mitogen-activated protein; ATP, adenosine triphosphate; CSP, circumsporozoite protein; MSP, merozoite surface protein; MESA, mature parasite-infected-erythrocyte surface antigen; MAEBL, membrane antigen erythrocyte binding-like.

Cluster	Genes	Num. known genes	Num. adjac. pairs	Stage	Probability of overlap by chance	Representative genes	Location
1	108	45 41%	12 (2)	Sporozoite	1.9×10^{-11} *Rifin* 3.9×10^{-8} Antigenic variation	25 different *rifin* genes 4 *stevor* genes *P. bergei* pbs36-related protein MAEBL, CSP, *var*(common) Sporozoite surface protein 2	Subtelomeric MAP kinase
2	152	32 21%	10 (4.3)	Sporozoite Gametocyte		Serine/threonine-protein kinase Gene 11-1, chitinase Mei-2 homolog	Central
3	218	28 12%	20 (8.72)	Gametocyte		Transmission blocking antigen 230 Transmission blocking antigen 48/45 Meiotic recombination protein dmc1 25 kD ookinate surface antigen *P. falciparum* gametocyte antigen 377	Central
4	95	32 33%	18 (1.6)	Ring Schizont Merozoite	2.9×10^{-2} Lipid metabolism 4.8×10^{-2} Fatty-acid metabolism	Early transcribed membrane protein 2.1, 11.1, 11.2 Erythrocyte-binding antigen 175, skeleton-binding protein, ring-expressed surface antigen	Subtelomeric
5	109	42 38%	3 (2)	Ring Early Trophozoite		ATP-dependent DNA helicase, putative RNA helicase, putative, *P. falciparum* erythrocyte membrane protein 1 (PfEMP1)	Central
6	167	110 65%	6 (5.2)	Ring Trophozoite	3.2×10^{-4} Macromolecule biosynthesis 6.2×10^{-4} Hexose metabolism	Ribosomal subunit Elongation factor Eukaryotic translation initiation factor RNA helicase (3), PfEMP3	Central
7	93	49 52%	2 (2)	Ring Trophozoite	7.4×10^{-5} Cytosolic ribosome	40*S* ribosomal subunits 60*S* ribosomal subunits Ribosomal protein L Falcipain-2, chloroquine resistance transporter	Central
8	125	54 43%	9 (3)	Trophozoite Gametocyte		MESA, *3 stevors, 2 rifins* Multidrug resistance protein 2	Central
9	120	34 41%	8 (3)	Trophozoite Gametocyte		Cell cycle control protein cwf15 Cytochrome c1 Cytochrome c2	Central
10	226	94 41%	13 (9.4)	Trophozoite Gametocyte	6.9×10^{-5} DNA replication 3.7×10^{-3} Mitotic cell cycle	6 DNA replication factor C subunits DNA polymerase delta 2 Proliferating cell nuclear antigens POM1	Central
11	110	58 52%	2 (2)	Sporozoite Trophozoite Schizont Gametocyte	4.2×10^{-10} Proteasome endopeptidase 1.97×10^{-9} 26S proteosome 2.15×10^{-4} 20S core proteosome	15 proteasome genes DNA repair protein rhp16 Facilysin	Central
12	238	95 39%	16 (10)	Trophozoite	6.8×10^{-4} Mitotic cell cycle 7.2×10^{-3} Mitochondrion	Cytochrome c oxidase DNA polymerase Origin recognition complex DNA primase Histones	
13	181	60 33%	21 (6.12)	Schizont Gametocyte		Cysteine proteases (5 SERAs) Cytoadherence linked protein 9 5 DNA replication licensing factors	Central
14	163	42 25%	11 (3.8)	Sporozoite Schizont Gametocyte	2.0×10^{-3} Actin filament-based process	Cyclin g-associated kinase, putative Merozoite surface protein 4 Myosin a and d Actin, actin depolymerizing factor	Central
15	130	38 29%	17 (3.12)	Schizont	2.7×10^{-8} Cell invasion 1.4×10^{-4} Apical complex 1.3×10^{-3} Rhoptry	3 Cytoadherence linked asexual proteins Erythrocyte binding antigen 140 MSP1,2,3,5,6,7, MSP8-like Merozoite capping protein-1 3 Reticulocyte binding proteins 4 Rhoptry proteins Subtilisin-like protease 2	Central

pressed genes described as proteasome endopeptidases would be found in the same cluster of 110 genes by chance is 4.4×10^{-10}. Although it is difficult to define the complete set of known invasion-specific genes, the probability that eight of the nine expressed genes with gene ontology annotations of "cell invasion" would be found in one cluster of 130 genes by chance is 2.7×10^{-8}. Thus, we conclude that a gene's expression profile can give clues about the gene's cellular role and that uncharacterized genes within a cluster are likely to be involved in cellular processes represented by one or more characterized genes within the cluster. A brief description of each cluster is given below.

Cluster 1 contains 108 genes that were highly expressed in sporozoites. Among these genes, 41 code for characterized proteins. This group includes most, and potentially all, known sporozoite-specific genes, such as the sporozoite surface protein 2 and the circumsporozoite (CS) protein (*25*), which are each induced ~800-fold in sporozoites relative to the erythrocyte stages. PFA0380w, a *P. falciparum* homolog of a *P. bergei* gene known to be highly expressed in sporozoites, was also found in this cluster (*25*), as well as the *P. falciparum* homolog of UIS1, a recently identified *P. yoelli* sporozoite-specific gene (*26*). The gene PF11_0486, which codes for MAEBL, a unique member of the erythrocyte binding ligand (EBL) family, is up-regulated in sporozoites and is also found in this cluster (*27–29*). Other genes found in this cluster may also be involved in the sporozoite's invasion of the liver cell and conceivably could represent targets for functional disruption of the liver stage invasion process, either by small molecule interactions or by vaccination.

Cluster 2 contains 152 genes, of which 79% encode hypothetical proteins. These genes were expressed at low levels in the erythrocytic and sporozoite stages and were highly expressed at the gametocyte stage. These genes generally code for proteins involved in cell-cycle regulation, such as mitogen-activated protein kinase 1 (PF14_0294), or for sexual differentiation, such as mitogen-activated protein kinase 2 (*30*, *31*). Additional DNA binding related proteins or protein kinases and phosphatases suggested that hypothetical genes in this cluster might be involved in the regulation of the cell cycle during the gametocytogenesis.

Cluster 3 contains 218 genes (88% hypotheticals) that were highly expressed only in gametocytes and includes known gametocyte-specific genes such as those coding for transmission-blocking target antigens PF13_0247 and PFB0405w (at least 23-fold and 4-fold induction, respectively), as well as the ookinete surface antigen (*32*) (Pfs25, 130-fold change) and PfNEK1 (PFL0080c, 130-fold change), a protein proposed to be involved in sexual differentiation (*33*). This cluster contains the one *Plasmodium* gene known to be involved in meiosis, the meiotic recombination dmc1–like protein, MAL8P1.76, and is consistent with gametocytes preparing for the sexual cycle. Clusters 2 and 3 together contain gametocyte-specific genes that might be targets for transmission-blocking interventions.

Cluster 4 contains 95 genes mainly highly expressed at the early ring stage and the late schizont stage of the erythrocytic cell cycle. Cluster 4 includes genes coding for proteins that are needed to parasitize the erythrocyte, such as ring-infected erythrocyte antigen (RESA, PFA0110w); early transcribed membrane proteins 11.1 and 11.2 (PF11_0039 and PF11_0040, respectively); and the newest member of PfSEB or *etramps* family, PFB0120w (*23*, *24*). PF07_0128, a member of the *ebl* family, was also found in this cluster, together with a number of genes involved in lipid metabolism (MAL6P1.62, PFI0695c, and PF14_0761; 2.9E-2) whose products may play a role in the establishment of the parasitophorous vacuolar membrane following invasion. Hypothetical proteins in this cluster may play a role in late cell invasion processes and early establishment of the erythrocytic environment.

Clusters 5, 6, and 7 contain genes that were expressed in ring and trophozoite stages, with expression levels that decline at the schizont stage. Differences between these clusters were mainly due to the breadth of the expression patterns throughout the erythrocytic cycle. Genes that belong to these clusters code for proteins involved in protein synthesis, such as translation initiation factors and ribosomal proteins. These clusters also contain most members of the glycolytic pathway. Most of these genes are only expressed at low levels in the merozoite, gametocyte, and sporozoite stages. Because hemoglobin degradation is specific to the trophozoite stage, most genes involved in hemoglobin degradation or the parasite's sensitivity to chloroquine are likely to be found in these clusters. Indeed, cluster 7 contains falcipain 2, a plasmepsin, and pfCRT, a gene involved in chloroquine resistance (*34*).

Clusters 8 to 13 contain genes with expression focused on the trophozoite stage, with slight time shifts and broad variation in expression patterns. The known genes are mainly involved in cell-cycle regulation and progression. Clusters 10, 12, and 13 contain almost all differentially expressed genes involved in DNA replication and are divided into different clusters according to whether they have roles in initiation or elongation. All six genes encoding components of DNA replication factor C are in cluster 10, and five of the six genes encoding DNA replication licensing factors, which, in yeast, are transcribed just after mitosis, are in cluster 13; five of the six differentially expressed SERA antigens (cysteine proteases) are in cluster 13. Members of these clusters may represent potential targets for drugs focused on disruption of this highly replicating stage of the parasite.

Cluster 14 contains genes that were highly expressed at the schizont stage as well as the sporozoite and gametocyte stages. This cluster contains a number of genes coding for actin and myosin and could be essential for apicomplexan host cell invasion and gliding motility. Indeed, myosin A (a motor protein), which localizes beneath the plasma membrane of the invasive stage, belongs to this cluster, as well as the newly identified myosin tail-interacting domain protein (MTIP) (*35*).

Cluster 15 contains 130 genes that were highly transcribed at the schizont stage, and 65% of the 37 characterized genes code for proteins that have been described as having a role in cell invasion ($P = 2.7 \times 10^{-8}$). Members include six of the eight merozoite surface proteins (MSPs), all proteins annotated as functioning in the rhoptries (organelles that play a role in invasion), and members of the protein family involved in erythrocyte binding. Most genes that have known roles in invasion or are under evaluation as blood-stage vaccines reside in this cluster, suggesting that additional candidate vaccine antigens could come from the uncharacterized members of this cluster.

For each of the clusters, expression profiles of all the gene members within the group were then reclustered with hierarchical methods to allow researchers to focus on any subset of genes that clearly form tighter local clusters. All the *k*-means clusters and the hierarchical clustering results for $k = 10$, 15, 20, 25, and 30 are available from our Web site (http://carrier.gnf.org/publications/CellCycle).

***Rifin*, *stevor*, and *var* gene expression.** The ability of the *P. falciparum* parasite to alter the complement of antigenically variant proteins that are presented on the infected cell surface facilitates chronic infection and transmission. Such proteins are encoded by at least three multigene families [*rifin*, *stevor*, and *var*, reviewed in (*36*)]. We observed different life cycle–dependent transcription patterns for the different families: *Rifins* were mostly up-regulated in sporozoites (cluster 1), as previously reported (*7*), although several showed trophozoite-specific expression patterns (cluster 8) consistent with the detection of RIFIN protein on the surface of the infected red cell (*37*). Differentially expressed *stevors* were also found in cluster 8, consistent with the protein's immunolocalization to the Mauer's cleft of the parasitized red cell (*38*), or in cluster 1 (*7*), whereas most *var* genes showed a slight early ring and sporozoite pattern. These patterns are unlikely to be the result of cross-hybridization between members of these multigene families. First, we confirmed by reverse transcription polymerase chain reaction (RT-PCR) stage-specific transcription of 16 *rifin*, *stevor*, or *var* genes. Second, using Basic Local Alignment Search Tool-Nucleotide (BLASTN), we showed that each 25-nucleotide probe maps to a unique location in the genome. Finally, genomic DNA hybridization shows that the "expression level" for these genes is no higher than for other genes, as would be expected if cross-hybridization were occurring (fig. S6).

Although the roles of *rifin* and *stevor*-encoded proteins are not well understood, *var*-encoded proteins (PfEMP1s) mediate adherence to host endothelial receptors, resulting in sequestration of infected red cells (*39*). Though there are 59 *var* genes in the parasite genome, evidence suggests that only one *var* gene is highly transcribed in any one parasite during the trophozoite stage of the life cycle when sequestration occurs (*40*). *Var* gene switching, which is regulated at the level of transcription, occurs at frequencies estimated at between 2 and 18% per generation (*41*). Because *var* gene expression and cytoadherence is associated with disease severity, the mechanism by which *var* gene expression is controlled has been an area of intense investigation. Sequencing of the *P. falciparum* genome (*5*) revealed that many *var* genes located in centromere proximal regions of the chromosome share a well-conserved upstream region, whereas *var* genes located in the subtelomeric regions have an alternate, conserved 5′ region, leading to the testable hypothesis that the two classes of *var* genes would show different regulatory patterns within the life cycle. Indeed, 34 of the 35 subtelomeric *var* genes having more than 1 probe on the array showed low, constitutive levels of transcription in the life cycle, with small relative increases mostly in the very early ring or sporozoite stages. However, only one subtelomeric *var*, MAL6P1.1, exhibited differential expression using the criteria described above (fold change of 4, cluster 14). One other subtelomeric gene, *var*(common) or PFE1640w, which shows similarity to *var* genes but lacks the second exon and is known to be broadly expressed, is found in cluster 1 (*42*). On the other hand, although some of the central *var* genes were also transcriptionally silent or constitutively expressed at low levels, several of the centrally located *var* genes showed dramatic changes in transcript levels across the life cycle. In particular, three physically adjacent central *var* genes on the left arm of chromosome 4 showed large fold inductions and expression in trophozoite-stage parasites in both cell cycles. In particular, PFD0625c (cluster 5) is induced 40-fold and 150-fold in cell cycles S and T, respectively. One other central *var* gene was considered differentially expressed (MAL7p1.50, cluster 14), but the fold change and absolute levels were low. These data suggest that the chromosome 4 loci could be the source of PfEMP1 expression in 3D7 trophozoites and that allelic exclusion could be occurring at the transcriptional level between these three or four loci in our 3D7 parasites. Although subtelomeric *var* genes appear to be relatively transcriptionally silent in these experiments, they clearly may still act as reservoirs of antigenic variation that can be spliced into the active central loci through recombination (*43*) in a mechanism analogous to mating-type switching in yeast.

Correlation between the transcriptome and the proteome. Based on its presumed function in mediating cytoadherence in the red blood cell, the detection of PfEMP1 protein in mosquito-stage parasites surprised many researchers when the analysis of the proteome was published (*7*). However, for the *var* genes on the array, we observed signal above background in sporozoites for 22 of the 23 genes whose cognate proteins had been detected by mass spectrometry in sporozoites. In fact, only 72 of the 1039 proteins detected by proteomic analysis of merozoites or trophozoites were not detected in our analysis (*7*), and 35 of these were single-spectra hits. However, the fact that levels of both PfEMP1 protein (measured by sequence coverage) and *var* gene transcription in sporozoites are low and expression is widespread suggests that neither transcription nor translation may be tightly regulated in the parasite. Alternatively, the low levels may just reflect the leading edge of a dramatic up-regulation that might occur when the sporozoite encounters a human hepatocyte (*41*).

Chromosomal organization and coexpressed genes. An analysis of the chromosomal organization of the differentially expressed genes grouped according to our cluster classification indicates that many highly expressed genes with a function in the erythrocyte-remodeling process, such as the genes in cluster 4, are located in regions near the ends of the chromosomes. The bulk of 55 subtelomeric genes that are differentially expressed (of 336 total) are in cluster 1 (25 genes, $P = 2.18 \times 10^{-22}$), 4 (11), 8 (6), or 15 (4). Conversely, genes having roles in growth and maintenance of the malaria parasite, such as those in cluster 6, were restricted to the central regions of chromosomes (fig. S5B). This observation highlights the fact that transcriptional regulation may be controlled in part at the chromosome level. The fact that a higher-than-expected number of genes from the same cluster were found to be adjacent to one another (Table 1) is additional evidence for chromatin-dependent regulation of transcription. This is particularly true for clusters with characteristic functions, such as cluster 1, in which genes are transcribed at the sporozoite stage; cluster 4, containing genes involved in the early establishment of the erythrocytic environment; and cluster 15, the "invasive cluster."

Altogether, these data provide a detailed description of the transcriptional events that occur throughout the life cycle of the malaria parasite. In addition, we have shown that the identification of genes with similar expression patterns can facilitate the determination of possible function of hypothetical proteins and, as a consequence, can provide clues about the potential cellular roles of more than 1000 hypothetical proteins. In some cases these data can lead to reinterpretation of existing data. For example, *clag9* was originally thought to play a role in cytoadherence based on its expression early in the cell cycle (*44*). Our data showed that *clag9* is expressed later in the cell cycle (cluster 13), suggesting an alternative role for the gene product. Indeed, analysis of its sequence revealed homology to proteins that are localized to the rhoptries. Immunolocalization experiments have now shown that CLAG9 protein is found in the rhoptries (*45*) as predicted by these data.

The reproducibility of the gene clustering data derived from independent synchronization methods and the high statistical significance of most clusters by ontology analysis emphasize the potential of using a whole *P. falciparum* genome microarray, as well as sophisticated algorithms and visualization tools, for examining *Plasmodium* biology. Of course, these data are not to be taken as perfect: Gene models may be incorrect; all probes may have failed for some small proportion of the genes, giving an incorrectly low expression value; and the choice of $k = 15$ clusters may have resulted in over- or underfragmentation of clusters. Indeed, proteins were detected for 34 genes that we designate as "not expressed." In addition, for very highly expressed genes, it is likely that we have exceeded the linear range of the array, resulting in an underestimation of transcript abundance. Furthermore, the genome sequences on the array were derived from an isolate, 3D7, that has been maintained in culture for generations, and our RNA samples were also derived from parasites that had been maintained in continuous culture for generations; therefore, our expression patterns may not reflect those that exist in parasites replicating human hosts. Obtaining the sequence of a wild *P. falciparum* isolate and obtaining expression profiles of parasites isolated directly from humans will improve the biological relevance of these studies. However, for most genes, both the cluster assignments and the expression profiles are likely to be quite reliable and accurate on the basis of our evaluation of characterized genes within the data sets. Our data disagreed with published results in only a few cases, and in such cases the published results may have involved strains or experiments that were substantially different from our own, allowing little to be concluded from any comparison. In addition, errors in the published gene annotations may also exist, thus confounding our ability to evaluate our data using ontology analysis. An increasing number of conditions, such as chemical stress or drug treatment, which result in sets of genes being up-regulated or down-regulated in a coordinated fashion, will provide additional data and allow the clusters to be refined, as may reevaluation of gene models and gene annotations. Ultimately, these data in combination with sequence data should be regarded as starting points for those interested in validating some of the thousands of proteins identified in the *P. falciparum* genome sequencing projects as new drug or vaccine targets.

References and Notes

1. J. G. Breman, A. Egan, G. T. Keusch, *Am. J. Trop. Med. Hyg.* **64**, iv (2001).
2. S. Bowman *et al.*, *Nature* **400**, 532 (1999).
3. M. J. Gardner *et al.*, *Science* **282**, 1126 (1998).
4. M. J. Gardner *et al.*, *Nature* **419**, 531 (2002).
5. N. Hall *et al.*, *Nature* **419**, 527 (2002).
6. R. W. Hyman *et al.*, *Nature* **419**, 534 (2002).
7. L. Florens *et al.*, *Nature* **419**, 520 (2002).
8. E. Lasonder *et al.*, *Nature* **419**, 537 (2002).
9. Y. Wu, C. D. Sifri, H. H. Lei, X. Z. Su, T. E. Wellems, *Proc. Nat.l Acad. Sci. U.S.A.* **92**, 973. (1995).
10. M. R. van Dijk, A. P. Waters, C. J. Janse, *Science* **268**, 1358 (1995).
11. B. S. Crabb, A. F. Cowman, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7289 (1996).
12. D. A. Fidock *et al.*, *Eur. J. Immunol.* **27**, 2502 (1997).
13. D. J. Lockhart, E. A. Winzeler, *Nature* **405**, 827 (2000).
14. P. K. Rathod *et al.*, *Trends Parasitol.* **18**, 39 (2002).
15. Z. Bozdech *et al.*, *Genome Biol.* **4**, R9 (2003).
16. K. G. Le Roch, Y. Zhou, S. Batalov, E. A. Winzeler, *Am. J. Trop. Med. Hyg.* **67**, 233 (2002).
17. Materials and methods are available as supporting material on *Science* Online.
18. C. Lambros, J. P. Vanderberg, *J. Parasitol.* **65**, 418 (1979).
19. J. D. Haynes, J. K. Moch, *Methods Mol. Med.* **72**, 489 (2002).
20. Y. Zhou, R. Abagyan, *BMC Bioinformat.* **3**, 3 (2002).
21. K. G. Le Roch *et al.*, in preparation.
22. I. W. Sherman, *Microbiol. Rev.* **43**, 453 (1979).
23. C. Birago *et al.*, *Mol. Biochem. Parasitol.* **126**, 209 (2003).
24. T. Spielmann, D. J. Fergusen, H. P. Beck, *Mol. Biol. Cell.* **14**, 1529 (2003).
25. S. H. Kappe *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9895 (2001).
26. K. Matuschewski *et al.*, *J. Biol. Chem.* **277**, 41948 (2002).
27. J. H. Adams, P. L. Blair, O. Kaneko, D. S. Peterson, *Trends Parasitol.* **17**, 297 (2001).
28. P. L. Blair, S. H. Kappe, J. E. Maciel, B. Balu, J. H. Adams, *Mol. Biochem. Parasitol.* **122**, 35 (2002).
29. M. Ghai, S. Dutta, T. Hall, D. Freilich, C. F. Ockenhouse, *Mol. Biochem. Parasitol.* **123**, 35 (2002).
30. C. M. Doerig *et al.*, *Gene* **177**, 1 (1996).
31. D. Dorin *et al.*, *J. Biol. Chem.* **274**, 29912 (1999).
32. M. Tachibana, T. Tsuboi, T. J. Templeton, O. Kaneko, M. Torii, *Mol. Biochem. Parasitol.* **113**, 341 (2001).
33. D. Dorin *et al.*, *Eur. J. Biochem.* **268**, 2600 (2001).
34. D. A. Fidock *et al.*, *Mol. Cell* **6**, 861 (2000).
35. L. W. Bergman *et al.*, *J. Cell Sci.* **116**, 39 (2003).
36. S. Kyes, P. Horrocks, C. Newbold, *Annu. Rev. Microbiol.* **55**, 673 (2001).
37. S. A. Kyes, J. A. Rowe, N. Kriek, C. I. Newbold, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9333 (1999).
38. M. Kaviratne, S. M. Khan, W. Jarra, P. R. Preiser, *Eukaryot. Cell* **1**, 926 (2002).
39. X. Z. Su *et al.*, *Cell* **82**, 89 (1995).
40. Q. Chen *et al.*, *Nature* **394**, 392 (1998).
41. M. L. Gatton, J. M. Peters, E. V. Fowler, Q. Cheng, *Trends Parasitol.* **19**, 202 (2003).
42. G. Winter *et al.*, *Mol. Biochem. Parasitol.* **127**, 179 (2003).
43. L. H. Freitas-Junior *et al.*, *Nature* **407**, 1018 (2000).
44. K. R. Trenholme *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4029 (2000).
45. Ling *et al.*, in preparation.

46. We thank H. Vial and J. Vinetz for critical reading of the manuscript and for helpful suggestions. This work was supported by a grant to E.A.W. from The Ellison Foundation. The sequence data were produced by the *Plasmodium falciparum* Sequencing Groups at the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/databases/P.falciparum_sequences), The Institute for Genomic Research (www.tigr.org), and the Stanford DNA Sequencing and Technology Center (www-sequence.stanford.edu/group/malaria). The International Malaria Genome Sequencing Project was supported by awards from the Burroughs Wellcome Fund, the Wellcome Trust Fund, NIAID, and the U.S. Department of Defense. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, the Department of Defense, or the U.S. government.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1087025/DC1
Materials and Methods
Figures S1 to S6
Tables S1 and S2
References

20 May 2003; accepted 23 July 2003
Published online 31 July 2003;
10.1126/science.1087025
Include this information when citing this paper.

REPORTS

Kinematic Evidence for an Old Stellar Halo in the Large Magellanic Cloud

Dante Minniti,[1]* Jura Borissova,[1] Marina Rejkuba,[2] David R. Alves,[3] Kem H. Cook,[4] Kenneth C. Freeman[5]

The oldest and most metal-poor Milky Way stars form a kinematically hot halo, which motivates the two major formation scenarios for our galaxy: extended hierarchical accretion and rapid collapse. RR Lyrae stars are excellent tracers of old and metal-poor populations. We measured the kinematics of 43 RR Lyrae stars in the inner regions of the nearby Large Magellanic Cloud (LMC) galaxy. The velocity dispersion equals 53 ± 10 kilometers per second, which indicates that a kinematically hot metal-poor old halo also exists in the LMC. This result suggests that our galaxy and smaller late-type galaxies such as the LMC have similar early formation histories.

In the Milky Way, the old metal-poor objects, such as globular clusters and RR Lyrae stars, define an almost spherical halo population (*1*–*5*). Models of halo formation by accretion (*6*) indicate that these old objects formed in small satellite galaxies that were subsequently accreted by the Galaxy, whereas dissipational collapse models (*7*) indicate that the halo formed rapidly before the disk collapsed. If these models apply to small galaxies, we would expect them to show a halo population defined by its oldest objects (*8*). At a distance of 50 kpc (*9*), the ideal laboratory to test this is the LMC, which is 10 times fainter than our galaxy. The oldest LMC globular clusters appear to lie in a flat rotating disk whose velocity dispersion is 24 km/s (*10*, *11*). This disk suggests that the LMC has indeed no kinematical halo of old metal-poor objects and that therefore the formation of the LMC proceeded without a halo phase. We measured the kinematics of field RR Lyrae stars in the LMC, which are known to be among the oldest and most metal-poor objects in this galaxy (*5*). This sample is a by-product of the MACHO (massive compact halo objects) microlensing project (*12*–*14*), which provided photometry for about 8000 RR Lyrae stars in a 10-square-degree region around the bar of the LMC (*15*).

The observations were acquired with the multislit Focal Reducer/Low Dispersion Spectrograph 1 (FORS1) at the European Southern Observatory (ESO) Very Large Telescope (VLT) Unit Telescope 1 (UT1) during the nights of 10 and 11 January 2003.

[1] Department of Astronomy, Pontificia Universidad Católica, Aveñida Vicuña Mackenna 4860, Casilla 306, Santiago 22, Chile. [2] European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching bei München, Germany. [3] Columbia Astrophysics Laboratory, 550 West 120th Street, New York, NY 10027, USA. [4] Institute of Geophysics and Planetary Physics, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. [5] Research School of Astronomy and Astrophysics, Australian National University, Mount Stromlo Observatory, Canberra, ACT, Australia.

*To whom correspondence should be addressed. E-mail: dante@astro.puc.cl