AD_____

Award Number: DAMD17-98-2-8005

TITLE: Malaria Genome Sequencing Project

PRINCIPAL INVESTIGATOR: Malcolm J. Gardner, Ph.D.

CONTRACTING ORGANIZATION: The Institute for Genomic Research
Rockville, Maryland 20850

REPORT DATE: January 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2001	Annual (17 Dec 99 - 16 Dec 00)

4. TITLE AND SUBTITLE
Malaria Genome Sequencing Project

5. FUNDING NUMBERS
DAMD17-98-2-8005

6. AUTHOR(S)
Malcolm J. Gardner, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Institute for Genomic Research
Rockville, Maryland 20850

E-MAIL:
Gardner@tigr.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims have been added: **Specific Aim 4**, sequencing if *P. yoelii* to 3X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 3X coverage. To date, we have published the first complete sequence of a malarial chromosome (chromosome 2 (7)), completed the random phase sequencing and closed most of the gaps in 3 other large chromosomes totaling 7.2 Mb, and have initiated functional genomics studies using glass slide micorarrays to characterize the expression of genes from chromosome s 2, 3, 10,11,12 and 14 throughout the erythrocytic cycle. We also collaborated in the construction of a two-enzyme optical restriction map of the entire *P. falciparum* genome (14), and are continuing to further develop the GlimmerM gene finding software developed in year 1. In addition, we have completed sequencing of the rodent malaria *P. yoelii* to 5X coverage and this year will begin sequencing of *P. vivax* to 3X coverage. A subcontract was awarded to the Scripps Research Institute to use proteomics techniques to identify sporozoite proteins in *P. falciparum* and *P. yoelii* sporozoites, leading to the idenfication of over 300 sporozoite proteins. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

14. SUBJECT TERMS
Breast Cancer, Plasmodium falciparum, Plasmodium yoelii, Plasmodium vivax, malaria, genome, microarray, proteomics

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Front Cover 1
SF298 2
Table of Contents 3
Introduction 4
Body 4
Sequencing of *P. falciparum* chromosome 14 (Specific Aims 1, 2, 3) 6
Sequencing of chromosomes 10 and 11 (Specific Aim 1) 8
Annotation 9
Microarray studies (added to Specific Aim 1) 10
Sequencing of *P. yoelii* and *P. vivax* to 3X coverage (Specific Aims 4, 5) 12
Proteomics studies 13
Key Research Accomplishments 14
Reportable Outcomes 14
Conclusions 15
References 16
Appendix A: Personnel Receiving Salary from this Cooperative Agreement 18
Appendix B: Progress report on subcontract to DAMD17-98-2-8005 19

Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many tropical areas of the world, and also affects many individuals and military forces that visit these areas. In 1994 the World Health Organization estimated that there were 300-500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably (17). These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control (4). Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism (2), and establishes an excellent starting point for this process. In 1995, an international consortium including the National Institutes of Health, the Wellcome Trust, the Burroughs Wellcome Fund, and the US Department of Defense was formed (Malaria Genome Sequencing Project) to finance and coordinate genome sequencing of the human malaria parasite *Plasmodium falciparum*, and later, a second, yet to be determined, species of *Plasmodium*. Another major goal of the consortium was to foster close collaboration between members of the consortium and other agencies such as the World Health Organization, so that the knowledge generated by the Project could be rapidly applied to basic research and antimalarial drug and vaccine development programs worldwide.

Body

This report describes progress in the Malaria Genome Sequencing Project achieved by The Institute for Genomic Research and the Malaria Program, Naval Medical Research Center, under Cooperative Research Agreement DAMD17-98-2-8005, over the 12 month period from Dec. '99 to Dec '00. The Specific Aims of the work supported by this agreement are listed below. Specific Aims 1-3 were contained in the original Cooperative Agreement. Specific Aims 4-5 were added to the Cooperative Agreement through modifications.

1. Determine the sequence of 3.5 megabases of the *P. falciparum* genome (clone 3D7):

a) Construct small-insert shotgun libraries (1-2 kb inserts) of chromosomal DNA isolated from preparative pulsed-field gels.

b) Sequence a sufficiently large number of randomly selected clones from a shotgun library to provide 10-fold coverage of the selected chromosome.

c) Construct P1 artificial chromosome (PAC) libraries (inserts up to 20 kb) of chromosomal DNA isolated from preparative pulsed-field gels.

d) If necessary, generate additional STS markers for the chromosome by i) mapping unique-sequence contigs derived from assembly of the random sequences to chromosome, ii) mapping end-sequences from chromosome-specific PAC clones to YACs.

e) Use TIGR Assembler to assemble random sequence fragments, and order contigs by comparison to the STS markers on each chromosome.

f) Close any remaining gaps in the chromosome sequence by PCR and primer-walking using *P. falciparum* genomic DNA or the YAC, BAC, or PAC clones from each chromosome as templates.

2. Analyze and annotate the genome sequence:

a) employ a variety of computer techniques to predict gene structures and relate them to known proteins by similarity searches against databases; identify untranslated features such as tRNA genes, rRNA genes, insertion sequences and repetitive elements; determine potential regulatory sequences and ribosome binding sites; use these data to identify metabolic pathways in *P. falciparum.*

3. Establish a publicly-accessible *P. falciparum* genome database and submit sequences to GenBank.

4. Perform whole genome shotgun sequencing of the rodent malaria parasite *Plasmodium yoelii* to 3X coverage, assemble into contigs, annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.

5. Perform whole genome shotgun sequencing of the human malaria parasite *Plasmodium vivax* to 3X coverage, assemble the contigs, annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.

We are pleased to report that excellent progress has been made towards achievement of these goals. In previous annual reports we announced the publication in *Science* of the first complete sequence of a malarial chromosome (chromosome 2) (7). In addition, we reported on work done by the TIGR/NMRC team and collaborators to provide new tools and resources for the Malaria Genome Project, including development of a *Plasmodium* gene finding program, GlimmerM (19), and introduction of optical restriction mapping technology for rapid mapping of whole *Plasmodium* chromosomes (11, 14). We also reported that the random phase of sequencing of 3 additional *P. falciparum* chromosomes had been completed, and that closure of gaps in these chromosomes was underway. We also began using microarray technology to

examine the expression of all genes from chromosomes 2 and 3 of *Plasmodium*. And finally, to facilitate community access to the sequence data, a *P. falciparum* genome web site was established at TIGR which contains all of the chromosome 2 sequence data and annotation, as well as preliminary sequences for the 3 other chromosomes currently being sequenced (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html).

In the past year our major efforts were directed to closure of chromosomes 10, 11 and 14. completion of these chromosomes is expected this year. In response to input from the malaria research community and the funders of the malaria genome consortium we also released preliminary annotation of these chromosomes on the TIGR web site (http://www.tigr.org/tdb/edb2/pfa1/htmls/). We also completed sequencing of the rodent malaria parasite *Plasmodium* yoelii to 3X coverage, and placed preliminary annotation of this genome on the TIGR web site (http://www.tigr.org/tdb/edb2/pya1/htmls/). We continued our close interactions with the Malaria Program at NMRC and collaborated with them in preparation of glass slide microarrays containing PCR fragments from almost all genes from chromosomes 2, 3, 10, 11, 12 , and 14. Experiments to profile the expression of these genes through the erythrocytic stage of the life cycle are underway. And through a subcontract to Dr. John Yates at the Scripps Institute, we also assisted NMRC in their pilot project to apply the techniques of proteomics towards the identification of novel antigens in parasite (sporozoite) extracts. Finally, we are continually reviewing with NMRC further steps that can be taken to more rapidly apply *Plasmodium* genomics, functional genomics, and proteomics to problems of vaccine development for malaria.

Sequencing of *P. falciparum* chromosome 14 (Specific Aims 1, 2, 3)

Sequencing of chromosome 14 (3.4 Mb) is being funded primarily by a grant from the Burroughs Wellcome Fund; funds from this collaborative agreement are being used to accelerate the sequencing, assist in closure and annotation, develop microarrays for chromosome 14, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. In previous years we described the isolation of chromosome 14 DNA, preparation of shotgun libraries , random sequencing, assembly, and progress in gap closure. Efforts this year have been directed towards closure of gaps in the chromosome, production and public release of preliminary annotation, and preparations for the final annotation.

The gap closure process began in December 1998. The procedures being used to close gaps are basically the same as those used previously on the chromosome 2 project(7), namely 1) use of GROUPER software to identify groups (contigs linked by shotgun clones), physical gaps and sequence gaps; 2) editing of contigs ends to remove untrimmed vector sequence, low quality sequence data, and chimeric clones that prevent merging of contigs; 3) resequencing of missing mates and short sequences at contig ends; 4) sequencing of shotgun clones spanning sequence gaps using primers at the ends of the gaps; 5) PCR with genomic DNA to span physical gaps; and 6) use of the transposon insertion method to close very AT-rich gaps. In practice, GROUPER is run on the set of contigs produced by an assembly and some or all of steps 1-6 are performed until no further progress is possible. Another assembly is then performed with the

edited contigs, new sequences (*e.g.* primers walks and missing mates), and unassembled sequences left over from the previous assembly. The new assembly will incorporate new sequences such as primer walks produced during closure, sequences edited during closure, and other sequences that did not get merged into the previous assembly, thereby providing new starting points for additional work. This process is repeated until the sequence is closed.

As noted above, due to cross-contamination of the chromosome 14 DNA with sheared nuclear DNA, up to 20% of the sequence data was derived from chromosomes other than chromosome 14. In order to focus the closure efforts on chromosome 14 contigs, chromosome 14 markers were used to identify which contigs and groups of contigs are from chromosome 14. With chromosome 2 about 30 markers were available (1 marker per 30 kb). In contrast, for chromosome 14 there are 98 STS markers derived from YACs (provided by Alister Craig) plus an additional 101 SSLP markers (21), providing a marker about every 17-20 kb. The higher density of markers has allowed identification of more chromosome 14 contigs and has simplified the gap closure process. In addition, with funding provided by the BWF, David Schwartz's group completed a 2-enzyme optical restriction map of the *P. falciparum* genome (14). We have used the optical map and the chromosome 14 markers to determine the order of contig groups on the chromosome, which reduced the number of PCR reactions required for closure of the physical gaps.

In last year's report we outlined plans for closure of the remaining physical and sequence gaps, most of which were expected to be very AT-rich and difficult to close. As shown in Table 1, the closure work has proceeded steadily, such that there are only 4 physical gaps and 2 sequence gaps to be closed. Over the past year, sequence gaps were closed by a combination of primer walks across plasmid templates spanning the gaps, and for very AT-rich sequence gaps, by sequencing transposon insertion libraries of clones spanning the gaps. This work is very labor intensive; closure of a single sequence gap can require the production of one or more transposon libraries, sequencing of 96 subclones, editing of the forward and reverse sequences from each transposon library, and reassembly of the sequences to close the gap. In addition, for some gaps this process needs to be repeated if the sequence quality in some AT-rich regions is not adequate. For closure of physical gaps, primers were predicted and synthesized at the ends of contigs adjoining the gaps, and PCR reactions were performed from genomic DNA to isolate fragments spanning the gaps. These PCR products were then sequenced directly to provide sequences spanning the gaps. However, many PCR products from physical gaps were very AT-rich and could not be sequenced directly. These AT-rich PCR products were cloned into plasmid vectors and used to make transposon libraries. Multiple transposon containing subclones were then isolated and sequenced, the sequences were edited, and the sequences were assembled to close the gaps. Again, closure of a physical gap frequently required several attempts, and in many cases multiple PCR products using nested primers were used to ensure that the a physical gap had been spanned properly. We expect to have closed all gaps by March 15, 2001.

Table 1. Progress in gap closure of *P. falciparum* chromosome 14 (only last 2 cycles of closure shown).

	7/99	12/00
Sequences	76,406	83,546
Largest contig (kb)	164	737
Total groups	ND[a]	5
Cum. Length (Mb)	ND	3.44
Physical gaps	ND	4
Sequence gaps	~ 64	2

[a] ND, not determined.

Once all gaps have been closed, the sequence will be evaluated with the program check_coverage to ensure that a) all regions of the assembly are covered by at least two shotgun clones, and b) that every base pair in the sequence has been sequenced in both directions with one chemistry, or in one direction with two chemistries. These criteria ensure that the sequence has been assembled correctly and validate individual base calls. The latter criterion is often satisfied by performing 10% of the sequence reactions with dye-primer chemistry. However, given the frequency of sequence artifacts in AT-rich regions observed with the dye-primer chemistry, this may not be appropriate for *P. falciparum*. As we discovered with chromosome 2, inclusion of sequences containing artifacts in an assembly inhibits contig formation and increases the number of sequence gaps in the assembly and the effort required to close them. Consequently, all chromosome 14 sequencing were done with dye-terminator chemistry. and late in the closure phase the coverage status of the assembly will be assessed. Regions with one-direction coverage will be identified, and additional dye-terminator reactions selected from the database will be performed to convert as many as possible to two-direction coverage. Regions with one-direction coverage that remain and which have unresolved sequence ambiguities will then be re-sequenced with dye-primer chemistry. This process will ensure that the coverage criteria are satisfied and minimize potential assembly problems arising from use of dye-primer chemistry. Finally, the sequence will be edited using the program TIGR_Editor, which displays all gel reads and electropherograms for each base in the sequence. Discrepancies will be noted and additional sequencing reactions will be performed to resolve ambiguities. As a last step to confirm colinearity of the assembled sequence and genomic DNA, restriction maps predicted from the sequence will be compared with the chromosome 14 optical restriction maps described above.. Once the chromosome 14 sequence has been completed, the sequence will be annotated for publication (see the annotation section).

Sequencing of chromosomes 10 and 11 (Specific Aim 1)

Chromosomes 10 and 11, which together constitute 16% of the genome (1.7 and 2.0 Mb, respectively), are being sequenced primarily with funding provided by the National Institute for Allergy and Infectious Diseases (L.M. Cummings is the Principal Investigator). Funds from this collaborative agreement are being used to accelerate the sequencing, assist in closure and

annotation, develop microarrays for these chromosomes, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. The random phase for chromosomes 11 and 10 was completed in mid- and late- 1999, respectively, and these chromosomes are now in closure. The closure procedure for these chromosomes is very similar to that used for chromosomes 2 and 14 (see chromosome 14 section above), and will take advantage of any technical improvements that are produced. One major difference in the closure process, however, is that many fewer microsatellite markers are available for these chromosomes(20, 21), making physical gap closure more difficult by reducing the number of contigs that can be accurately ordered on the chromosome. Consequently, Dr. Cummings is collaborating with Dr. X Su of the Laboratory of Parasitic Diseases, NIAID, in production of additional microsatellite markers for these chromosomes. Currently, there are only 10 sequence gaps and 1 physical gap in chromosome 11; completion of this chromosome is expected by June 1, 2001. Chromosome 10, which was started later, is not so far along. Groups of contigs totaling 1.4 Mb have been ordered along the chromosome and contain 4 physical gaps and 11 sequence gaps. Another region of about 350 kb is in several groups of contigs and have not yet been ordered along the chromosome. Multiplex or combinatorial PCR (22) with primers from the ends of these groups using genomic DNA as template will be performed in order to localize these contigs along the chromosome, and then sequence and physical gaps will be closed as described above. Raw sequence reads, preliminary contigs, and preliminary annotation of chromosomes 10 and 11 have been released on the TIGR web and will be updated periodically as closure proceeds (http://www.tigr.org/tdb/edb2/pfa1/htmls/pfa1.shtml).

Annotation

Over the past year there have been major improvements and changes to the databases, software, and web displays for annotation of eukaryotic genomes at TIGR. Much of this was based on work done to develop an annotation system for the *Arabidopsis* genome at TIGR, and we have modified the *Arabidopsis* system for use with *Plasmodium* falciparum. Some development work is still underway, but we used the new system to prepare an automated, preliminary annotation of chromosomes 10, 11, and 14. This preliminary annotation was released on a new set of *Plasmodium* falciparum pages on the TIGR web site in October 2000 (http://www.tigr.org/tdb/edb2/pfa1/htmls/). This web site represents a vast improvement over the previous web site that TIGR had released in Oct. '98 upon the publication of *P. falciparum* chromosome 2. We are currently in the process of moving the chromosome 2 data into the new annotation database so as to have all 4 *P. falciparum* chromosomes sequenced at TIGR accessible in a single database and web site.

The automated annotation system performs the following tasks. Upon deposit of a contig in the Pfa1 annotation database, a set of scripts collectively called eukaryotic genome control, or egc, launches the gene finder GlimmerM. GlimmerM, software developed at TIGR by Ela Pertea and Steven Salzberg specifically for this project, predicts gene models in the sequence. The gene models are then searched against protein databases and the results are parsed and recorded in the database. The gene models are also searched against HMM's representing protein families in the PFAM(1) and TIGRFAM(8) databases. A program called AUTOBYOB then scans the results of the database and HMM searches and attempts to provide a tentative functional assignment for the

predicted protein, e.g. "aldolase." Genes without obvious functional assignments are given the assignment "hypothetical protein." The tentative assignments are recorded in the database.

The preliminary annotation web site provides a variety of tools for searching of the database and visualization of the annotation. For example, the tentative gene assignments can be searched by keyword; a list of genes matching the keyword is displayed, and clicking on links provided allows the user to view annotation associated with each gene. Users can also use a contig viewer that displays a map of a contig and the genes contained in the contig. In addition, the new site allows users to download complete contig sequences as well as the nucleotide and predicted amino acid sequences of individual genes. Further enhancements to the display of preliminary annotation will be added to the site over the next year.

In addition to building a system to provide preliminary annotation of unfinished contigs from chromosomes 10, 11, and 14, we have made preparations for the final annotation of the completed chromosome sequences. The foundation of the system is the new Pfa1 relational database that was used to prepare and display the automated preliminary annotation. A new annotation tool called AnnotationStation, developed for TIGR by Neomorphic, Inc. for the *Arabidopsis* project, presents a graphical display of the chromosome sequence and the results of all analyses performed on the sequence (e.g. GlimmerM gene models, results of BLAST searches against protein and nucleotide databases, predicted splice sites, etc.). The annotator can view all of this information and make working models of each gene. The working models are edited by the annotator, and are then saved in the pfa1 database. The egc system then produces a predicted protein of each working gene model and searches the protein sequence against a protein database using BLASTP, and against PFAMs and TIGRFAMs, and also predicts protein structural features such as signal peptides and transmembrane domains. After these analyses have been performed, a program called auto_byob attempts to automatically assign a common name (e.g. aldolase) and a functional role to the gene. An annotator can view all of the evidence associated with a particular gene using the Submit program, and decisions regarding the gene identification made by the annotator are recorded in the database. As described above for preliminary annotation, the database then provides the information required for construction of the outside web displays, and for submissions to GenBank and to PlasmoDB.

Microarray studies (added to Specific Aim 1)

In last year's report we described our first efforts to add functional genomics studies to this *P. falciparum* genome sequencing project. The aim of these functional genomic studies is to provide a more complete view of *Plasmodium* biology by determining gene expression information for all *Plasmodium* genes that are identified through the genome sequencing effort. We chose to use glass slide microarrays for this work (3). Microarrays can be used to examine the expression patterns of thousands of genes simultaneously from two or more RNA samples. These RNA samples may be derived from parasites grown under different growth conditions, or from different life cycle stages, in order to determine the complement of genes that may be differentially expressed under varying conditions. Most of this work is being conducted by CDR Carucci and Dr. Adam Witney at the NMRC. TIGR is working closely with the NMRC group to provide preliminary sequence data and annotation, primers for amplification of gene sequences,

and other materials that are required for this project. The arrays are also printed at TIGR using our arraying robots and software. We coordinate closely on protocols and software development. Dr. John Quackenbush, director of TIGR's microarray facility, is also participating in these studies.

In pilot studies conducted at NMRC to evaluate this technology, PCR products representing virtually all genes from chromosomes 2 and 3 were prepared and arrayed on glass slides using TIGR's Molecular Dynamics Arrayer robot. Oligonucleotides corresponding to at least one exon in the identified chr2/3 genes were designed, synthesized and used to amplify the specific region from genomic DNA. An automated primer design program was written (using Primer3 from Whitehead Institute) to produce primers with Tm=55∞C (+/- 2 degrees C) and amplicons between 250 and 750 bp. Of the 550 primers pairs designed, over 90% gave a good single PCR product of the expected size. These PCR products were arrayed on poly-L lysine slides in 50% DMSO using a robotic microarrayer (through agreement with TIGR). All amplicons were spotted in triplicate on each slide. The DNA was fixed to the surface of the glass slide by UV cross linking at 90 mJ. *P. falciparum* parasites (clone 3D7) were cultivated using standard methods and were synchronized in a temperature cycling incubator. Blood stage total RNA was prepared by Trizol and the cDNA labeled with either Cy3-dUTP or Cy5-dUTP using oligo dT priming. The labeled cDNA was applied to the surface of the microarray slide and allowed to hybridize at 65 degrees C overnight in aqueous solution containing non-homologous sheared DNA. After a stringent wash, the slides were scanned using a ScanArray 3000 laser scanner and the resultant images analyzed using Imagene 3.0 to determine the signal, background fluorescence intensity at each spot. The text results were stored in a customized Sybase relational database and accessed via a Web interface..

In optimization experiments, review of the pattern of hybridization in genes where more than open reading frame was used, revealed that the further the PCR product was predicted from the 3' end of the gene, the lower the signal intensity at that spot. This result was expected, however there were cases where the signal intensity dropped even as close as 2500 bp from the 3' end of the gene. Possibly explanations include: long untranslated 3' tail, sequence-specific reduced cDNA labeling efficiency, overall poor cDNA synthesis, or incorrectly predicted ORFs. In order to maximize the signal intensities from subsequent arrays, primer design will include biasing the primer design to the 3' end of the gene model. Further, other cDNA labeling techniques will be assessed including the use of Cy-labeled duct (vice Cy-labeled dump) and indirect labeling of pre-synthesized cDNA (to avoid the inhibitory effect of incorporating Cy-labeled dyes during the cDNA synthesis).

Analysis of the data revealed clear examples of differential gene expression during the erythrocytic cycle, which encouraged us to proceed with construction of new microarrays containing all of the genes identified in completed chromosome sequences and the chromosomes being sequenced at TIGR and Stanford University (chromosomes 10, 11, 12, and 14). These arrays will contain PCR products from approximately 30% of the parasite genome, and can be expanded later to include most parasite genes as sequences from other chromosomes become available. Expression studies being conducted by NMRC are focusing on the description of gene expression patterns during the erythrocytic cycle, and the effect of anti-malarial drugs on parasite gene expression. A manuscript on these studies is expected to be completed this year.

Sequencing of *P. yoelii* and *P. vivax* to 3X coverage (Specific Aims 4, 5)

The primary goal of the Malaria Genome Sequencing Project was to sequence the genome of *P. falciparum*. The random sequencing phase of all 14 chromosomes was completed last year, 2 chromosomes have been finished, annotated and published, and several other chromosomes are nearing completion. Thus virtually all *P. falciparum* genes have at least partial sequences in the databases, and malariaologists have access to most *P. falciparum* genes. Even today, with only 2 chromosomes completed, this sampling of recent articles citing the genome data produced by the consortium shows that the genome project has had a major effect on malaria research (5, 6, 9, 10, 12, 13, 15, 16, 18, 23-25).

A secondary goal of the malaria genome project was to sequence the genome of a another species of *Plasmodium*, and discussions as to which parasite should be chosen had generated lively discussions amongst the malaria community, with some groups favoring sequencing of the human malaria *P. vivax*, and others advocating sequencing one of the rodent malaria parasites that are used as model systems. The sequence of one or more species would be very useful for comparison to *P. falciparum*, perhaps enabling the identification of genes that may be involved in differences in life cycles and pathogenicity, for example. Genome sequence information from other *Plasmodium* species would also be helpful in annotation of *P. falciparum*, by assisting in identification of genes conserved across different species. Recent discussions at the semi-annual meetings of the malaria genome consortium may lead to efforts funded by the NIAID, the Burroughs Wellcome Fund, or the Wellcome Trust, to do partial sequencing of several rodent malaria genomes, which will provide useful sequence data to groups working on these different parasites at a reasonable cost.

In light of these events, and the reductions in sequencing costs achieved by the TIGR SeqCore through improvements in instrumentation and sequencing protocols, the TIGR/NMRC team discussed the expansion of our sequencing efforts to include *P. vivax* and *P. yoelii*. *P. vivax* is a major human malaria parasite, and *P. yoelii* is a rodent malaria parasite used as a model system for vaccine development by NMRC. By using a whole genome shotgun strategy and sequencing to 3X coverage , it is possible to assemble contigs covering about 90% of the *Plasmodium* genome. With the high gene density of *Plasmodium*, this is a relatively rapid and low-cost method to acquire partial or complete sequences of almost all parasite genes.

After discussions with NMRC we elected to proceed with sequencing of *P. yoelii* (under the supervision of Dr. Leda Cummings), and then to sequence *P. vivax*. *P. yoelii* was given priority over *P. vivax* for sequencing despite the greater importance of *P. vivax* as a human pathogen because we did not have *P. vivax* genomic DNA suitable for sequencing (but see below). Genomic shotgun libraries with 1.5 kb inserts were prepared in the plasmid vector pHOS1, and 270,332 sequencing reactions were performed, producing 221,415 sequences with an average length of 661 nt (sequencing success rate = 82%). This is equivalent to 5.85X coverage assuming a genome size of 25 Mb (this does not take into account possible contamination from mouse genomic DNA; we are investigating this now). The entire data set has not yet been assembled, but assemblies were performed at 1X, 2X, and 3X coverage; contigs and preliminary annotation of the 2X data were made available on the TIGR web site (http://www.tigr.org/tdb/edb2/pya1/htmls/). Analysis of the contigs obtained at 3X coverage indicated that 8,103 contigs were obtained with an average length of 2.9 kb, and a cumulative

length of 24.1 Mb. Over the next few months we will assemble the entire 5.8X coverage data set into contigs, perform a preliminary automated annotation, and post the data and annotation on the TIGR web site.

This year we will begin sequencing of the SalI strain of *P. vivax* to 3X coverage. CDR Carucci at NMRC will prepare *P. vivax* genomic DNA in collaboration with Niconor Obaldia at the Gorgas Memorial Institute in Panama . Genomic shotgun libraries will be prepared and sequenced at TIGR, and contigs and preliminary annotation will be posted on the TIGR web site as they become available.

Thus, within one year most of the *P. falciparum* genome will have been sequenced to completion and annotated, *P. yoelii* will have been sequenced to approx. 5-6X coverage, and assuming genomic DNA is available by June 2001, *P. vivax* will have been sequenced to 3X coverage. The malaria research community will have access to the complete genome sequence of the most deadly human malaria parasite (*P. falciparum*), a draft sequence of *P. vivax*, which is the second most important human malaria parasite, and a draft sequence of *P. yoelii*, a rodent malaria parasite used as a model system in malaria vaccine development.

Proteomics studies

A major goal of the malaria genome project is to identify antigens for vaccine development. Analysis of the genome sequence data can be used to identify potential antigens but does not by itself provide all of the information required for selection and prioritization of vaccine candidates. For example, the genome sequence itself does not specify at which point in the life cycle a gene is transcribed, or whether the protein product of a gene is actually present in the parasite. To gather information on gene expression patterns we initiated the microarray studies in collaboration with NMRC that are described above. To identify proteins present in various stages of the parasite life cycle, we have begun to use proteomics techniques to directly identify parasite proteins in cell lysates.

This work is being done by Dr. John Yates at the Scripps Research Institute, partly funded by a subcontract from TIGR under this cooperative agreement. Briefly, proteins in parasite lysates are digested with proteases and the resulting peptides are separated by high resolution liquid chromatography. The peptides are then injected into a tandem mass spectrometer. Spectra of each peptide are then matched against predicted spectra of the peptides predicted from the genome sequence. In this way peptides generated from cells lysates can be used to identify the proteins present in the cell lysate. Dr. Yates performed a series of such experiment using sporozoites of *P. falciparum* and *P. yoelii*, identifying 308 unique proteins from *P. falciparum*, and 37 unique proteins from *P. yoelii*. This represents a massive increase in the number of known sporozoite proteins, and indicates that the same techniques can be used to identify proteins present in other stages of the life cycle. When combined with information gleaned from the sequence data, such as predicted subcellular location, hydophillicity, and predicted T cell epitopes, the protein expression data will help to prioritize potential antigens for vaccine development.

A report prepared by Dr. Yates as required by the conditions of the subcontract is attached as Appendix A.

Key Research Accomplishments

1) Closure of chromosomes 10, 11, and 14 of *Plasmodium falciparum* was the major focus of work over the past year. All 3 chromosomes are now in the late stages of gap closure. Completion and annotation of these chromosomes is expected by December, 2001.

2) Preliminary sequence data and annotation for chromosomes 10, 11, and 14 was released on the TIGR web site (http://www.tigr.org/tdb/edb/pfdb/pfdb.html).

3) A new annotation database called Pfa1 was prepared and a semi-automated annotation pipeline was developed. This annotation system is now being enhanced in preparation for the final annotation of the chromosomes.

4) Shotgun sequencing of the rodent malaria parasite *P. yoelii* to 3X coverage was completed; preliminary contigs and annotation were released on a new TIGR web site http://www.tigr.org/tdb/edb2/pya1/htmls/).

5) Plans to shotgun sequence the *P. vivax* genome to 3X coverage were prepared in consultation with NMRC. Sequencing will begin once DNA has been provided to TIGR by NMRC.

6) Microarrays containing PCR fragments representing the genes identified on chromosomes 2, 3, 10, 11, and 14 have been prepared by NMRC and pilot studies to establish labeling, hybridization, and detection protocols were completed. Experiments to characterize gene expression in erythrocytic parasites are underway at NMRC.

7) Under a subcontract to Dr. John Yates at the Scripps Research Institute, liquid chromatography and tandem mass spectrometry was used to identify proteins expressed in sporozoites of *Plasmodium falciparum* and *Plasmodium yoelii*.

Reportable Outcomes

1) Web site. GlimmerM gene finder (http://www.tigr.org/softlab/glimmerm/)

2) Web site. Preliminary contigs and annotation for *P. falciparum* chromosomes 10, 11, and 14. (http://www.tigr.org/tdb/edb2/pfa1/htmls/).

3) Web site. Preliminary contigs and annotation for *P. yoelii* genome at 3X coverage. (http://www.tigr.org/tdb/edb2/pya1/htmls/).

4) M. J. Gardner, L. M. Cummings, Invited presentation: Sequencing of P. falciparum chromosomes 10, 11, and 14, American Society of Tropical Medicine and Hygiene, Washington, D.C. (1999).

5) M. J. Gardner, Invited presentation: Progress report on sequencing of P. falciparum chromosome 14, Malaria Genome consortium Meeting, Sanger Center, Hinxton, England. (2000).

6) L. M. Cummings, Invited presentation: Progress report on sequencing of P. falciparum chromosomes 10 and 11, Malaria Genome Sequencing Consortium, Sanger Centre, Hinxton, England (2000).

7) Patent application. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum and proteins of said chromosome useful in antimalarial vaccines and diagnostic reagents. Filed by NMRC. Note: this was reported under this section in last year's report, but we neglected to include DD Form 882. We include a copy of the patent application and DD Form 882.

8) Subcontract to Dr. John Yates, Scripps Research Institute. See Appendix A and DD 882.

Conclusions

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims have been added: **Specific Aim 4**, sequencing if *P. yoelii* to 3X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 3X coverage. To date, we have published the first complete sequence of a malarial chromosome (chromosome 2 (7)), completed the random phase sequencing and closed most of the gaps in 3 other large chromosomes totaling 7.2 Mb, and have initiated functional genomics studies using glass slide microarrays to characterize the expression of genes from chromosome s 2, 3, 10, 11, 12 and 14 throughout the erythrocytic cycle. We also collaborated in the construction of a two-enzyme optical restriction map of the entire *P. falciparum* genome (14), and are continuing to further develop the GlimmerM gene finding software developed in year 1. In addition, we have completed sequencing of the rodent malaria *P. yoelii* to 5X coverage and this year will begin sequencing of *P. vivax* to 3X coverage. A subcontract was awarded to the Scripps Research Institute to use proteomics techniques to identify sporozoite proteins in *P. falciparum* and *P. yoelii* sporozoites, leading to the identification of over 300 sporozoite proteins. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

References

1. **Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, and F. Servant** 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites Nucleic Acids Res. **29:**37-40.
2. **Bloom, B. R.** 1995. A microbial minimalist Nature. **378:**236.
3. **Brown, P. O., and D. Botstein** 1999. Exploring the new world of the genome with DNA microarrays Nat Genet. **21:**33-7.
4. **Butler, D., J. Maurice, and C. O'Brien** 1997. Briefing malaria Nature. **386:**535-540.
5. **Carter, N. S., C. B. Mamoun, W. Liu, E. O. Silva, S. M. Landfear, D. E. Goldberg, and B. Ullman** 2000. Isolation and functional characterization of the PfNT1 nucleoside transporter gene from *Plasmodium falciparum* J Biol Chem. **275:**10683-91.
6. **Gardner, M. J.** 1999. The genome of the malaria parasite Current Opinion in Genetics and Development. **9:**704-708.
7. **Gardner, M. J., H. Tettelin, D. J. Carucci, L. M. Cummings, L. Aravind, E. V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pedersen, K. Shen, J. Jing, D. C. Schwartz, M. Pertea, S. Salzberg, L. Zhou, G. G. Sutton, R. L. Clayton, O. White, H. O. Smith, C. M. Fraser, M. D. Adams, J. C. Venter, and S. L. Hoffman** 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum* Science. **282:**1126-1132.
8. **Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White** 2001. TIGRFAMs: a protein family resource for the functional identification of proteins Nucleic Acids Res. **29:**41-3.
9. **Hayward, R. E., J. L. Derisi, S. Alfadhli, D. C. Kaslow, P. O. Brown, and P. K. Rathod** 2000. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria Mol Microbiol. **35:**6-14.
10. **Horrocks, P., K. Dechering, and M. Lanzer** 1998. Control of gene expression in *Plasmodium falciparum* Mol Biochem Parasitol. **95:**171-81.
11. **Jing, J., C. Aston, L. Zhongwu, D. J. Carucci, M. J. Gardner, J. C. Venter, and D. C. Schwartz** 1999. Optical mapping of *Plasmodium falciparum* chromosome 2 Genome Research. **9:**175-181.
12. **Jomaa, H., J. Wiesner, S. Sanderbrand, B. Altincicek, C. Weidemeyer, M. Hintz, T. r. I, M. Eberl, J. Zeidler, H. K. Lichtenthaler, D. Soldati, and E. Beck** 1999. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs Science. **285:**1573-1576.
13. **Kyes, S. A., J. A. Rowe, N. Kriek, and C. I. Newbold** 1999. Rifins: A second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum* Proc Natl Acad Sci U S A. **96:**9333-8.
14. **Lai, Z., J. Jing, C. Aston, V. Clarke, J. Apodaca, E. T. Dimlanta, D. J. Carucci, M. J. Gardner, B. Mishra, T. Anantharaman, S. Paxia, S. L. Hoffman, J. C. Venter, E.**

J. Huff, and D. C. Schwartz 1999. A shotgun optical map of the entire *Plasmodium falciparum* genome Nature Genetics. **23:**309-313.

15. **Le Roch, K., C. Sestier, D. Dorin, N. Waters, B. Kappes, D. Chakrabarti, L. Meijer, and C. Doerig** 2000. Activation of a *Plasmodium falciparum* cdc2-related kinase by heterologous p25 and cyclin H. Functional characterization of a *P. falciparum* cyclin homologue J Biol Chem. **275:**8952-8.
16. **Meinnel, T.** 2000. Peptide deformylase of eukaryotic protists: a target for new antiparasitic agents? Parasitol Today. **16:**165-8.
17. **Organization, W. H.** 1997. World malaria situation in 1994: population at risk Weekly Epidemiological Record. **72:**269-276.
18. **Ridley, R. G.** 1999. Planting the seeds of new antimalarial drugs Science. **285:**1502-1503.
19. **Salzberg, S. L., M. Pertea, A. Delcher, M. J. Gardner, and H. Tettelin** 1999. Interpolated Markov models for eukaryotic gene finding Genomics. **59:**24-31.
20. **Su, X., M. T. Ferdig, Y. Huang, C. Q. Huynh, A. Liu, J. You, J. C. Wootton, and T. E. Wellems** 1999. A Genetic Map and Recombination Parameters of the Human Malaria Parasite *Plasmodium falciparum* Science. **286:**1351-1353.
21. **Su, X. Z., and T. E. Wellems** 1999. *Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR Exp Parasitol. **91:**367-9.
22. **Tettelin, H., D. Radune, S. Kasif, H. Khouri, and S. L. Salzberg** 1999. Optimized Multiplex PCR: Efficiently Closing a Whole-Genome Shotgun Sequencing Project Genomics. **62:**500-507.
23. **Vinetz, J. M., J. G. Valenzuela, C. A. Specht, L. Aravind, R. C. Langer, J. M. Ribeiro, and D. C. Kaslow** 2000. Chitinases of the avian malaria parasite *Plasmodium gallinaceum*, a class of enzymes necessary for parasite invasion of the mosquito midgut J Biol Chem. **275:**10331-41.
24. **Wellems, T. E., X. Su, M. Ferdig, and D. A. Fidock** 1999. Genome projects, genetic analysis, and the changing landscape of malaria research Current Opinions in Microbiology. **2:**415-9.
25. **Woodrow, C. J., J. I. Penny, and S. Krishna** 1999. Intraerythrocytic *Plasmodium falciparum* expresses a high affinity facilitative hexose transporter J Biol Chem. **274:**7272-7.

Appendix A: Personnel Receiving Salary from this Cooperative Agreement

AHN, SUSIE
ANTHONY, ROBERT
BAILEY, MARSHA
BENTON, JONATHAN
BERA, JAYATI
BERRY, KRISTI
BORKOWSKI, NICOLE T.
BOWMAN, CHERYL
BRENNER, MICHAEL I.
BUCHOFF, JEFFREY
BURGESS,SHANDRECA
BURR, PATRICK
CALDWELL, LAUREN
CARTY, HEATHER
CHEN, MINGHUA
CIECKO, ANNE
COVARRUBIAS, MIGUEL
CRONIN, LISA
CUMMINGS, LEDA
CURTISS, RAHIM
DINTERMAN, SHELLY
DRAGOI, IOANA
DUGUE, NATALIE REDDIX
ELLIOTT, DAVID
FUJII, CLAIRE
GANGESTAD, DONNA
GANSBERGER, KRISTEN
GARDNER, MALCOLM
GARRETT, MINA
GEBREGEORGIS, ELIZABETH
GILL, JOHN
GLADNEY, EMILY
GRIMES, ELIZABETH
HANCE, MARK E.
HANSEN, CHERYL
HEREFORD, NATALIE
HILL, JESSICA
HOLMES, MICHAEL
IMPRAIM, MARJORIE
JACKSON, JACQUELINE
JARRAHI, BEHNAM
JENKINS, CHELTON
JENKINS, JENNIFER
JIANG, LINGXIA
JONES, KRISTINE
KALB, ERICA
KANG, KATHERINE
KURUSHKO, ALENA
LARKIN, CHRISTOPHER
LEE, KATHERINE
LEE, PERVIS C.
LEVITSKAIA, IRINA
LEWIS, MATTHEW
LYNN, JEFFREY
MAHURKAR, ANUP
MASON, TANYA
MILITSCHER, JENNIFER
MOAZZEZ, AZITA
PAI, GRACE H.
PARKSEY,DEBBIE
PARVIZI, BABAK
PETROGRADSKAYA, INNA
RADUNE (BUSHMAN), DIANA
RIGGS, FLORENCE
RIZZO, MICHAEL
ROMERO, CLAUDIA
ROONEY, TIMOTHY
RUCH, KAREN
RYABTSEVA, TAMARA
SCANLAN, DAVID
SELLERS, PATRICK
SENTER, JACQUELINE
SHALLOM, SHAMIRA
SHVARTSBEYN, ALLA
SMIRNOVA, TATYANA
STEWART, AMY
SUH, BERNARD
TALLON, LUKE J.
TRAN, BAO
TRAN, KEVIN
TSITRIN, TAMARA
UTTERBACK, TERESA
VAN AKEN, SUSAN
VON ARX, ANNA
WANLESS, DAVID
WEAVER, BRUCE
WILLIAMS, MARY
ZHAO, YONGMEI
ZSCHOCHE, CHRISTINA

Appendix B: Progress report on subcontract to DAMD17-98-2-8005

A PROTEOMIC VIEW OF THE MALARIA PARASITE LIFE CYCLE

JOHN R. YATES

DEPARTMENT OF CELL BIOLOGY

THE SCRIPPS RESEARCH INSTITUTE

LA JOLLA, CA 92037

INTRODUCTION

The malaria parasite is a perfect example on how one genome can encode several proteomes responsible for very different cellular forms, capable of invading both vertebrate and invertebrate hosts. With the *Plasmodium falciparum* genome sequence reaching completion and sequencing projects of other *Plasmodium* species under way, knowing the gene and protein expression changes that occur during the life cycle is within reach and would greatly enhance the fight against malaria[1], by pinpointing potential targets for drugs and vaccine development.

Proteomics allows for differential-expression studies at the protein level. Analysis of a cell proteome requires the resolution of the proteins in a sample followed by the identification of the resolved proteins. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) followed by mass spectrometry (MS) is the most widely used method of protein resolution and identification[2, 3, 4]. However, this approach has two main flaws. On the one hand, it is time-consuming: each protein resolved on 2D-PAGE has to be extracted from the gel, digested with proteases and the resulting peptides identified by MS. On the other hand, integral membrane proteins, proteins with extreme in pI and molecular weight and low abundance proteins are poorly resolved on 2D-PAGE.

Multidimensional protein identification technology (MudPIT) is a novel non-gel-based proteomic technique, which combines on-line high-resolution liquid chromatography and tandem mass spectrometry[5]. In MudPIT, a peptide mixture is generated prior to the chromatography, which makes this method independent of the physico-chemical properties of the proteins to be identified. Peptides from a complex mixture are eluted in an iterative process from a biphasic microcapillary column directly into an electrospray ionization ion trap mass spectrometer. Tandem mass spectra (MS/MS), which contain fragmentation patterns specific to amino acid sequences, are generated from peptides after they elute into the mass spectrometer. MS/MS spectra are assigned to peptides found in a sequence database by the SEQUEST algorithm[6].

MudPIT proved successful in identifying about 1500 proteins expressed in *S. cerevisiae* grown to mid-log phase[7], which is the largest proteome analysis to date. Furthermore, proteins rarely seen in classical proteome

---

[1] Hoffman, S.L. Research (Genomics) is crucial to attacking Malaria. *Science* **290**, 1509 (2000).

[2] Hanash, S.M. Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. *Electrophoresis* **21**, 1202-9 (2000).

[3] Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837-46 (2000).

[4] Washburn, M.P. & Yates, J.R. 3rd. Analysis of the microbial proteome. *Curr Opin Microbiol* **3**, 292-297 (2000).

[5] Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., & Yates, J.R. 3rd Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**, 676-82 (1999).

[6] Eng, J.K., McCormack, A.L. & Yates, J.R. 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976-989 (1994).

[7] Washburn, M.P., Wolters, D., & Yates J.R. 3rd. Large- Scale Analysis of the Yeast Proteome via Multidimensional Protein Identification Technology *Nat Biotechnol* in press.

analyses using 2D-PAGE and MS were also detected and identified. MudPIT is therefore the technique of choice for rapid, large-scale and largely unbiased proteome analysis.

Our main goal is to comprehensively assess the proteomes of different cellular forms of the malaria parasite, in collaboration with NMRC, which is providing us with parasite cells at various stages of the life cycle (sporozoites, liver schizonts and blood stage parasites from both the human parasite, *Plasmodium falciparum* and the mouse parasite, *Plasmodium yoelii*).

METHODS

Lysing cells — Whole cell protein lysates from *P. falciparum* sporozoites were obtained using 2 different lysis procedures. The cells were either diluted 10 times in Lysis buffer (310 mM NaF, 3.45 mM $NaVO_3$, 50 mM Tris, 12 mM EDTA, 250 mM NaCl, 140 mM $Na_4P_2O_7$, pH 7.60) and sonicated 3 times 30sec, or diluted 10 times in 0.1 M Sodium Carbonate, pH 11.6 and incubated in ice for 1 hour. Soluble and insoluble protein fractions were separated by centrifugation at 15,000 rpm for 30 min. Both methods released the same amount of total proteins per 10^6 cells. Sodium carbonate extraction was more efficient at removing proteins as a soluble fraction, *i.e.* the insoluble fraction obtained from this method is likely to mostly contain integral membrane proteins. *P. yoelii* sporozoites, erythrocytes infected with *P. falciparum* or *P. yoelii* parasites and control red blood cells from human and mouse were lyzed by sodium carbonate extraction.

Generating peptides — The pH of the soluble fractions was adjusted to 8.5 using 1 M Ammonium Bicarbonate. Proteins in the soluble fractions were denatured in the presence of 8 M urea and disulfide bonds were hydrolyzed by TCEP (5 mM). The resulting free cysteines were further modified with 10 mM Iodoacetamide (IAM). Proteins were digested by Endoproteinase Lys-C (Sequencing grade, Roche Diagnostics), overnight at 37°C. After diluting the samples to 2 M Urea with 100 mM Ammonium Bicarbonate pH 8.5, $CaCl_2$ was added to a final concentration of 2 mM. Poroszyme Bulk Immobilized Trypsin (Applied Biosystems) was used to further digest the proteins, at 37°C, overnight, while shaking. Trypsin beads were discarded by spinning at 15,000 rpm for 10 min.

Pellets constituting the insoluble protein fractions were solubilized in 100µl 90% formic acid, 500mg/ml CNBr, overnight, in the dark. Saturated NH_4HCO_3 was added drop by drop to neutralize formic acid, until pH is ~8.5. Neutralized samples were lyophilized down to about 500µl. Solutions were brought up to 8M urea by adding solid urea. From this point forward, the

samples were dealt with the same way as soluble fractions, *i.e.* with TCEP, IAM, endoproteinsase Lys-C and trypsin treatments.

The peptide mixtures obtained from soluble and insoluble fractions were loaded onto SPEC-PLUS PTC 18 cartridges (Ansys) for concentration and buffer exchange. Peptides were eluted off with 100 to 200µl 90% Acetonitrile, 0.5% Acetic Acid. Peptide samples were lyophilized down to 3 to 5µl and finally diluted back to a final volume of 20µl in 5% Acetonitrile, 0.5% Acetic Acid. Amino acid analysis was carried out on aliquots from each sample to estimate the amount of peptide material. Peptide mixtures were stored at -80°C until analyzed by LC/MS/MS.

Resolving peptides and collecting MS/MS datasets — The peptide mixture from each sample was loaded onto a fused-silica microcapillary column (100_m i.d. x 365_m o.d., Polymicro Technologies), packed first with 8 to 9cm of 5_m C_{18} reverse phase particles (XDB-C18, Hewlett Packard), followed by 4 to 5cm of 5_m strong cation exchange material (Partisphere SCX, Whatman). Resins were washed with 100% methanol for at least 10min and equilibrated in Buffer A (5% ACN, 0.02% HFBA) for at least 30 minutes before loading the peptide sample.

The loaded microcapillary column was installed such as to spray directly into a Finnigan LCQ ion trap mass spectrometer equipped with an nano-LC electrospray ionization source. The flow rate was set to about 200-300nl/min controlled by a Quaternary Agilent 1100 series HPLC pump. Such a quaternary system allowed the use of 4 different elution buffers: Buffer A, Buffer B (80% ACN, 0.02% HFBA), Buffer C (250mM ammonium acetate, 5% ACN, 0.02% HFBA), Buffer D (500mM ammonium acetate, 5% ACN, 0.02% HFBA). A fully automated 12-step chromatography run was carried out on each soluble fraction sample and either 12-step or 7-step runs were used to resolve peptides from insoluble fractions. In such sequences of chromatographic steps, peptides were sequentially eluted from the SCX resin to the RP resin by increasing salt steps (increase in Buffer C concentration), followed by RP gradient (increase in Buffer B concentration) to elute the peptides from the RP resin directly into the mass spectrometer. The last chromatography step consisted in a high salt wash with 100% Buffer D.

Full MS spectra were recorded over a m/z range of 400 to 2000, on the eluting peptides. Tandem MS (MS/MS) spectra were sequentially generated on the first, second and third most intense ions selected from the full MS spectrum. The tandem mass acquisition settings were a default charge state of 2, a default isolation width of 3, collision energy of 35 and the minimum signal required was 10,000. Dynamic exclusion was enable for 10 minutes and a mass width of 3.

Matching MS/MS Spectra to Peptides — A *P. falciparum* database of 7,613 open-reading frames (ORFs) was assembled from the published sequences of chromosomes 2[8] and 3[9],

[8] Gardner, M.J., Tettelin, H., Carucci, D.J., Cummings, L.M., Aravind, L., Koonin, E.V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D.C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G.G., Clayton, R., White, O., Smith, H.O., Fraser, C.M., Adams, M.D., Venter, C.J., Hoffman, S.L. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, 282: 1126-1132 (1998)

[9] Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C.M., Craig, A., Davies, R.M., Devlin, K., Feltwell, T., Gentles, S., Gwilliam, R., Hamlin, N., Harris, D., Holroyd, S., Hornsby, T., Horrocks, P.,

and the partially annotated sequences of chromosomes 12 available from the Stanford DNA Sequence and Technology Center[10], chromosomes [10-11-14] provided to us by TIGR/NMRC[11], and chromosomes [1-4-5-6-7-8-9-13] downloaded from the Sanger Center web site[12]. A *P. yoelii* peptide database was obtained from TIGR/NMRC (Leda Cummings, personal communication) and contained 34,833 ORFs deduced from the 3x coverage of the *P. yoelii* DNA sequence. Finally, both *P. falciparum* and *P. yoelii* databases were complemented with a set of 165 known protein contaminants, such as trypsin, bovine serum albumin and keratins.

The SEQUEST algorithm[13] was used to match MS/MS spectra to peptides in the *P. falciparum* and *P. yoelii* sequence databases. Differential search options were specified in the SEQUEST search parameters. To account for cysteine carboxymethylation and methionine oxidation, every MS/MS dataset was searched with 57 Da and 16 Da optionally added to the average molecular weight of cysteine (103.1388 Da) and methionine (131.1926 Da) residues, respectively. Furthermore to account for the modifications that occur when CNBr cleaves after methionine residues[14], the MS/MS data resulting from insoluble samples treated with CNBr and formic acid also had to be independently analyzed two more times with SEQUEST. For each run, the differential search modification was engaged on methionine and set to either –30 or –48 accounting for methionine transformation by CNBr to Homoserine or Homoserine lactone, respectively.

The validity of peptide/spectrum matches was assessed using the SEQUEST-defined parameters[15] cross-correlation score (XCorr), Delta Cn value, Sp rank and relative ion proportion. The DTASelect package[16] was used to select and sort peptide/spectrum matches passing a set of those parameters. Were only retained peptide/spectrum matches displaying a DeltCn value of 0.12, a minimum ion proportion of 40%, a maximum Sp rank of 50, and minimum XCorr scores of 1.95 for +1 spectra, 2.5 for +2 spectra, and 4.2 for +3 spectra. Those were conservative selection criteria compared to the commonly described ones[17, 18]. Finally, every peptide/spectrum match passing this selection was visually confirmed. A confidence score was assigned to those matches depending on two main criteria: (i) any given MS/MS spectrum had to be clearly above the baseline noise and (ii) both *b* and *y* ion series should show continuity.

RESULTS

Jagels, K., Jassal, B., Kyes, S., McLean, J., Moule, S., Mungall, K., Murphy, L., Barrell. B.G. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature*, 400: 532-538 (1999)

[10] http://baggage.stanford.edu/group/malaria/download.html

[11] http://www.tigr.org/tdb/edb/pfdb/pfdb.html

[12] ftp.sanger.ac.uk/pub/pathogens/malaria2/unfinished_ORFS/orf_peptide_sequences

[13] Eng, J.K., McCormack, A.L. & Yates, J.R. 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976-989 (1994).

[14] Aitken, A., Geisow, M.J., Findlay, J.B.C., Holmes, C. & Yarwood, A. Peptide preparation and characterization. in *Protein Sequencing: a practical approach* (eds. Findlay, J.B.C. & Geisow, M.J.) 43-68 (IRL Press, New York, 1989).

[15] Tabb, D.L., Eng, J.K., & Yates, J.R. 3rd. Protein identification by SEQUEST

[16] Tabb, D.L.

[17] Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., & Yates, J.R. 3rd. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**, 676-82 (1999).

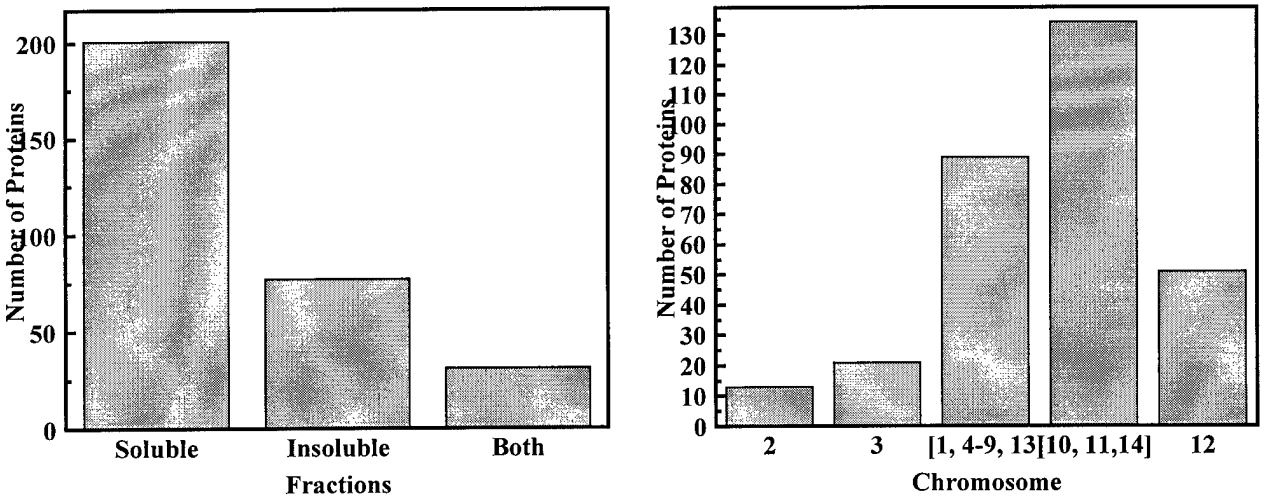
[18] Washburn, M.P., Wolters, D., & Yates J.R. 3rd. Large- Scale Analysis of the Yeast Proteome via Multidimensional Protein Identification Technology *Nat Biotechnol* in press.

***Plasmodium falciparum* sporozoites** — Four independent whole cell protein extracts were generated from *Plasmodium falciparum* sporozoites, each being further fractionated into soluble and insoluble fractions. All 8 peptide mixtures were analyzed by MudPIT. MS/MS spectra were matched to proteins in a *P. falciparum* database containing 7,613 ORFs. Peptide hits derived from SEQUEST searches were selected and sorted with the DTASelect package using a set of conservative parameters judging the quality of the peptide/spectrum matches. Every peptide/spectrum match passing those strict criteria was further assessed visually.

Compiling the peptide hits from all 8 samples using the CONTRAST package[19], 308 unique proteins were confidently identified from *P. falciparum* sporozoites. Two hundred proteins were found to be only present in soluble protein fractions, whereas 77 proteins were identified only from membrane protein fractions (Figure 1). Identified proteins were encoded by all chromosomes and represented between 4 to 6% of the ORFs predicted per chromosome (Figure 2).

As an example, among the proteins identified, were found the products of genes MP03001 on chromosome 3, 498.t00001 on chromosome 10 and 616.t00008 on chromosome 11. Those encode proteins known to be present in *P. falciparum* sporozoites, namely circumsporozoite protein, asparagine-rich protein and antigen 332.

***Plasmodium yoelii* sporozoites** — Three peptide mixtures were generated from two soluble fractions and one insoluble fraction isolated from *P. yoelii* sporozoites. The MS/MS datasets obtained were analyzed against a *P. yoelii* database containing DNA sequence translations from stop codon to stop codon, in all 3 open-reading frames on both strands,



resulting in 34,833 ORFs. Each peptide/spectrum match assigned by SEQUEST was assessed for confidence as described. The two soluble peptide mixtures and the insoluble peptide mixture contributed 30 and 9 proteins, respectively, to a final tally of 37 unique proteins. Lower quantities of peptide material were available for LC/MS/MS analysis compared to the studies on *P. falciparum* sporozoites. This mostly accounts for the disappointingly low number of proteins identified, to date, from *P. yoelii* parasites, but can be easily overcome.

[19] Tabb, D.L.

CONCLUSIONS

Those preliminary results prove MudPIT as an efficient tool to mine *P. falciparum* and *P. yoelii* genomes for information on protein expression. Further analyses of proteomes isolated from sporozoites and red blood cells infected with *P. falciparum* or *P. yoelii* — as well as the appropriate non-infected control proteomes from mosquitoes and red blood cells — are necessary and currently under way. Differential analysis of those proteomes should provide a clearer image of the protein expression changes occurring during the malaria parasite life cycle.