

Plasmodium post-genomics: an update

Daniel J. Carucci

Grand Challenges in Global Health Initiative, Foundation for the National Institutes of Health, 45 Center Drive (3AN-44), Bethesda, MD 20892-460, USA

The concept behind the first Molecular Approaches to Malaria meeting, held 1–5 February 2000 in Lorne, Australia, was ahead of its time; to convene a meeting of malaria researchers, database developers and genomics scientists, and to discuss how genomic sciences and their relevant disciplines could be applied to solve important problems in malaria research. The success of the second Molecular Approaches to Malaria meeting, held 1–5 February 2004 in the same place, together with the influence of genomics on malaria research, is testament to the vision that the organizers had at the first meeting. This review attempts to capture some of the current efforts in the post-genomics era of malaria research and highlights the approaches discussed at the Molecular Approaches to Malaria 2004 meeting.

At the time of the first Molecular Approaches to Malaria (MAM) meeting held 1–5 February 2000 in Lorne, Australia, the first two *Plasmodium falciparum* chromosomes had been published, establishing the feasibility of completing the entire genome sequence of this parasite. It took another two years before all 14 chromosomes would be published and the full impact of the completed genome on malaria research would be realized. How the data were to be presented to the research community in a manner that would allow the generation and answering of detailed questions resulted in the formation of a worldwide, accessible, relational database PlasmoDB (<http://www.plasmoDB.org>), which was initiated at the University of Pennsylvania, PA, USA. By February 2000, DNA microarray technology was applied to malaria research for the first time and large-scale transcriptional profiling of multiple stages of parasite development was initiated; the ability to carry out these transcriptional studies with both printed and Affymetrix arrays are now relatively facile. The foundation for high-throughput proteomic approaches, which was to a large degree developed for the study of yeast protein expression, was also just being described for *Plasmodium*.

Current state of malaria genomics

By the mid 1990s, there had been an explosion of genome sequencing projects, which fell on the heels of the publication of the genome of the first free-living organism, *Hemophilus influenzae* [1]. The sequencing of *P. falciparum* and its most important vector, *Anopheles gambiae*, were soon to follow [2]. A large-scale, shotgun method was used to sequence the first two chromosomes of *P. falciparum* [3] and the completion of these chromosome projects was by

no means trivial. With an AT content of ~80%, the *P. falciparum* genome presented substantial technical hurdles; DNA sequencing chemistries had to be altered and adapted, computer alignment algorithms were rewritten, novel gene-model prediction software was developed [4] and new approaches to verify the computer-generated sequence assembly were generated [5]. To tackle these challenges, an international genome consortium was created [6], resulting in the publication of the *P. falciparum* (clone 3D7) genome sequence by October 2002 [7].

The intention of the consortium was to generate a highly detailed genomic map of *P. falciparum* with a coverage range of between 12- and 15-fold redundancy. Because there is substantial synteny between *Plasmodium* spp., efforts were made to generate sequence data to around fivefold coverage from one of the rodent malaria parasites, *Plasmodium yoelii*, in parallel to the *P. falciparum* project. The completion of this second *Plasmodium* genome, which was at a fraction of the cost of sequencing the *P. falciparum* genome, is invaluable for comparative genomic studies and, in particular, for identifying potential vaccine and drug targets. The information derived from comparing genomic sequences can assist in the functional characterization of *Plasmodium* proteins of unknown function. For example, the *P. yoelii* genome contains 600–1000 copies of *vir* genes [8] located at the subtelomeric ends of the chromosomes, which is a characteristic of another large family of genes in *P. falciparum*, the *var* genes, that are believed to have a role in antigen variation and immune evasion. The *P. yoelii* sequence was published along with the complete genome sequence of *P. falciparum* in the same issue of *Nature* [9].

Several additional *Plasmodium* genome sequencing projects are now under way (see Z. Bozdech *et al.*, this issue), including: (i) *Plasmodium chabaudi*, a rodent malaria parasite; (ii) *Plasmodium knowlesi* and *Plasmodium reichenowi*, two non-human primate malaria parasites; (iii) *Plasmodium gallinaceum*, an avian malaria parasite; and (iv) *Plasmodium vivax*, a human malaria parasite. These projects are in various stages of completion and can be reviewed online at: <http://www.tigr.org/tdb/parasites/> and <http://www.sanger.ac.uk/Projects/Protozoa/>. Because there are recognized potential shortcomings in using a laboratory-maintained *P. falciparum* line, which is known to have substantial chromosomal deletions, there are also plans to sequence a *P. falciparum* isolate that has been obtained directly from a patient undergoing an exchange transfusion.

There are still many technical barriers facing malaria genomics despite the successes of previous *Plasmodium*

Corresponding author: Daniel J. Carucci (dcarucci@fnih.org).

Available online 8 October 2004

genome projects. Large-scale genome sequencing has become less costly, although significant investment of resources and expertise that is not generally available outside well-established genome centers is required. Large numbers of repeat sequences can make genome assembly challenging. Fortunately, the *P. falciparum* genome can be used as an assembly scaffold for other *Plasmodium* spp. because of the high level of synteny between *Plasmodium* genomes. However, automated annotation algorithms often differ greatly over predicted genes, hence greater emphasis is needed to improve gene-prediction methods and to verify those predictions experimentally. The research community can help here by providing feedback on experimentally verified sequences to the genome centers. It should be stressed that sequencing large-scale full-length expressed sequence tags (EST) will be extremely valuable [10] because they will inform the computer-generated gene-model predictions.

Databases and comparative genomics

PlasmoDB, the official database for malaria genome and functional genomics data, stores, organizes and displays a wide range of genomics data, and provides query tools to allow investigators to ask questions on the database. The PlasmoDB team at the University of Pennsylvania is continually adding new tools to the database and implementing novel methods of capturing a wide range of data types (such as gene and protein expression data). In addition, the PlasmoDB team is developing improved visualization tools to make the data more useful to the research community. GeneDB (<http://www.GeneDB.org>), another genomics database, focuses on displaying improved annotations (single 'best' gene models for each gene) and is contributing to improved annotations under the gene ontology consortium, which has undertaken the development of controlled vocabulary to describe gene products from a range of organisms (see A.E. Berry *et al.*, this issue).

Sequencing data from other apicomplexan species, such as *Toxoplasma gondii*, *Theileria* sp., *Babesia bovis*, *Eimeria tenella* and *Cryptosporidium* sp., will expand the repertoire of comparative genomic studies and will also provide insight into the complex biology across this phylum (<http://www.sanger.ac.uk/Projects/Protozoa/>). Comparative genomic analyses among a diverse range of *Plasmodium* spp. combined with functional genomics data will: (i) lead to a more complete understanding of parasite evolutionary biology; (ii) uncover mechanisms of host immune-evasion responses; (iii) identify factors that confer host and tissue specificity; and (iv) lead to targeted methods of interfering with parasite biology. For example, the ratio of non-synonymous (dN) to synonymous (dS) mutations (dN:dS ratio) might be greater in genes predicted to have secretory signal sequences or those present on cell membranes (M. Berriman, pers. commun.). This leads to the speculation that these genes are under selective immune pressure, and could be involved in host-parasite interactions or in avoidance of parasite clearance.

Proteomics

The characterization of large numbers of proteins in complex samples has revolutionized methods to determine

cellular and subcellular protein expression. This has also opened up new opportunities to complement and broaden the use of the *Plasmodium* genome sequence data. The general approach adopted in these large-scale proteomics studies is to first separate protease-generated peptide fragments in complex protein mixtures from parasites (soluble or insoluble fractions) using micro-liquid chromatography (micro-LC) or one-dimensional gel chromatography. The peptides are then measured using mass spectrometry to generate peptide fingerprints for each fragment, which are compared with predicted peptide fingerprints of proteins from genome databases [11,12]. The micro-LC and mass spectrometry method is capable of identifying thousands of proteins from complex protein mixtures [13–16] (for review, see Ref. [17]). Since the publication of two *P. falciparum* proteome studies [18,19], these approaches have been expanded to include proteomic studies of a range of parasite life stages, subcellular fractions, and even parasite-derived proteins embedded in the erythrocyte surface [20] or in association with specialized structures such as Maurer's clefts [21].

Proteomics approaches have also been used to identify proteins from asexual blood stages (including purified merozoites), gametocytes and sporogonic stages [18,19]. One approach to identify the subcellular location of proteins from a particular cellular compartment relies on an analysis and comparison of the relative abundance of proteins from purified organelles, host cellular compartments, or soluble and insoluble protein fractions, to predict which proteins are associated with each purified fraction. This approach can, for example, identify proteins on the parasite or host cell surface, which might be involved in host-parasite interactions. In addition, the characterization of proteins exposed to the host immune system that are associated with membrane structures and are potentially involved in transporter functions could lead to the putative identification of protein function. For example, proteomic studies of preparations enriched with parasite rhoptries are used to identify specific proteins associated with the rhoptries (T. Sam-Yellowe, pers. commun.). Proteomic studies could also be extended to other cellular compartments from various parasite life stages to gain more insight into the role of these specialized structures.

Merozoites and sporozoites from different *Plasmodium* spp. differ in the host tissue in which they invade and develop. A comparison of proteins expressed from these stages and from different species could help identify proteins that confer this host cell specificity. The merozoite proteomes of three *Plasmodium* spp. (*P. falciparum*, *Plasmodium knowlesi* and *Plasmodium berghei*) have been compared to determine a set of proteins that might be responsible for differential erythrocyte tropisms and thus host specificity (S.M. Khan *et al.*, Leids Universitair Medisch Centrum, Leiden, The Netherlands). This method could also help identify species-specific invasion processes, in addition to those that might be more generic in their role of invasion.

One of the key challenges in proteomics studies is obtaining a sufficient amount of parasites, both quantitatively and qualitatively, that minimizes host cell or

other protein contamination. *In vitro* cultures produce adequate numbers of blood-stage parasites for most proteomics studies, although purification of membranes from infected erythrocytes, or subcellular compartments (rhoptries, micronemes, food vacuoles) can be problematic. Sexual stages can be obtained from *P. berghei* and *Plasmodium gallinaceum*, particularly the ookinete stages, which are readily produced *in vitro* [22]. Sufficient numbers of sporozoites can be hand-dissected from ~1000 mosquito salivary glands and purified over a gradient column or by mechanical purification methods for certain functional genomic studies. However, the numbers here represent the limit of what is generally required based on current technologies; more recent and more-sensitive mass spectrometers are now available, which will work with smaller amounts of starting material. Difficulty in purifying sufficient quantities of a certain parasite stage of interest (e.g. hepatic stages) or of a particular subset (e.g. male and female gametes) could limit the use of proteomics in the study of these neglected stages.

One solution to the problem of parasite abundance is to use a selection method that enriches for a particular parasite stage, such as monoclonal antibody-coated magnetic beads that have been used to enrich sporozoite preparations [18]. Several parasite lines have been generated that express green fluorescent protein (GFP) and these lines offer the potential to generate highly purified preparations of parasites. *Plasmodium berghei* lines that express GFP under the control of promoters specific for expression within the male and female gametes have been sorted by flow cytometry to a high degree of purity. These pure preparations were then analyzed by mass spectrometry, revealing the expression of hundreds of male and female stage-specific proteins (S.M. Khan *et al.*, unpublished).

Other transgenic parasite lines expressing GFP under the control of different promoters will be needed to fully exploit this purification method. For instance, preliminary studies with GFP-expressing *P. berghei* and *P. yoelii* have the potential to sort infected hepatocytes either from *in vitro* infected hepatic cell lines or *in vivo* infected mouse liver (J. Williams and N. Martinez-Alier, pers. commun.).

Quantitative proteomics

Whereas gene chips and traditional quantitative molecular approaches (i.e. quantitative PCR or blotting) are readily available and can be used to quantify transcripts accurately, determining protein abundance on a large scale from small samples is not a routine method. Quantitative proteomics methods have been developed that can provide a measure of protein abundance between two samples. One method is based on efficient *in vivo* incorporation of a heavy isotope-labeled or a non-isotope amino acid during cell growth into the test and control samples, respectively. The relative peak sizes of the corresponding peptide fragments that are generated and measured by micro-LC mass spectrometry are then determined [23]. Because peptide mass and/or charge peak of the labeled sample will be shifted by the addition of the heavy amino acid, the abundance of the test sample

can be compared with that of the control by the differential area under the two peaks.

Recently, efforts have been made to apply this heavy-isotope amino acid labeling method to asexual-stage parasites grown in culture. However, it is difficult to select the appropriate isotope-labeled amino acid to add to the culture medium because *Plasmodium* derive a substantial amount of protein from hemoglobin digestion. Isoleucine is an ideal amino acid because it is required for parasite development, but is absent from human hemoglobin. Indeed, ¹⁵N-isoleucine has been efficiently incorporated into schizonts in *in vitro* culture and was successfully used to determine relative protein abundance in asexual stage parasite development [24]. This foundation work will probably result in expanded studies of altered quantitative patterns of protein expression and their correlation with gene chip transcript-expression studies.

Finally, proteomic methods are also being used to improve *P. falciparum* genome annotation by: (i) aiding gene-model predictions through the identification and validation of alternative splice junctions at the intron-exon boundaries [19]; and (ii) the identification of missing protein models by searching peptide fingerprints through six-frame-translated genome sequence (L. Florens, pers. commun.).

Structural biology of parasite proteins

Determining the three-dimensional structure of *Plasmodium* proteins has been particularly complicated because it is difficult to express sufficient levels of high-quality *Plasmodium* proteins for producing crystals as part of X-ray crystallographic structure determination or as protein for nuclear magnetic resonance (NMR) spectroscopy studies. This difficulty also applies to: (i) non-soluble proteins or proteins with significant hydrophobic domains; (ii) proteins with large non-globular domains; and (iii) glycosylated proteins. However, there are resources now available to make progress toward large-scale protein-structure determinations of *Plasmodium* proteins. The Structural Genomics Consortium (SGC; <http://www.thesgc.com/>) is an ambitious international consortium that has been established to determine three-dimensional structures of medically important proteins. In addition to their core efforts on human protein structure determination, one of the partners, SGC Toronto, is embarking on a project to purify several proteins from *Plasmodium* and other apicomplexans, to determine their three-dimensional structures and provide these as a public resource. Brannigan *et al.* (Structural Biology Laboratory, University of York, UK), as part of a functional genomics consortium, have also initiated approaches to develop methods for protein-target selection, protein production and crystallization from a series of soluble proteins that represent potential antimalarial drug targets. Although technically challenging, these efforts should contribute to information on protein function and potential new antimalarial drug targets (see also: <http://depts.washington.edu/sbpp>).

A glimpse into the future

Whereas the foundation for much of the efforts in genomics approaches to malaria research is in the completion of the genome sequence, as highlighted at the MAM 2000 meeting, the breadth of these additional approaches was apparent at MAM 2004. From outlining approaches to determine virulence loci in malaria parasites, to understanding parasite proteins involved in endothelial binding, and to using genomics approaches to provide additional evidence of the role merozoite surface protein 1 (MSP-1) plays in protective immunity, genomics-based malaria research is helping us to understand the host–parasite interaction in more detail. As new technologies that capitalize on the genome sequence are applied to malaria research, in addition to a consortia of biologists, physicists, computational biologists and others working on forming a more complete picture of parasite biology, we get closer to the development of interventions that can limit the mortality and morbidity caused by this important parasite.

References

- 1 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512
- 2 Holt, R.A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298, 129–149
- 3 Bowman, S. *et al.* (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* 400, 532–538
- 4 Salzberg, S.L. *et al.* (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics* 59, 24–31
- 5 Lai, Z. *et al.* (1999) A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat. Genet.* 23, 309–313
- 6 Hoffman, S.L. *et al.* (1997) Funding for the malaria genome sequencing project. *Nature* 387, 647
- 7 Gardner, M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511
- 8 del Portillo, H.A. *et al.* (2001) A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* 412, 839–842
- 9 Carlton, J.M. *et al.* (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419, 512–519
- 10 Watanabe, J. *et al.* (2004) Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, *Plasmodium* species. *Nucleic Acids Res.* 32 Database issue, D334–D338
- 11 Eng, J.K. *et al.* (1994) 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
- 12 Yates, J.R. 3rd *et al.* (1997) Direct analysis of protein mixtures by tandem mass spectrometry. *J. Protein Chem.* 16, 495–497
- 13 Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682
- 14 Washburn, M.P. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 19, 242–247
- 15 Wolters, D.A. *et al.* (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 73, 5683–5690
- 16 Koller, A. *et al.* (2002) Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11969–11974
- 17 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207
- 18 Florens, L. *et al.* (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419, 520–526
- 19 Lasonder, E. *et al.* (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 419, 537–542
- 20 Florens, L. *et al.* (2004) Proteomics approach reveals novel proteins on the surface of malaria-infected erythrocytes. *Mol. Biochem. Parasitol.* 135, 1–11
- 21 Sam-Yellowe, T.Y. *et al.* (2004) A *Plasmodium* gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. *Genome Res.* 14, 1052–1059
- 22 Janse, C.J. *et al.* (1985) *In vitro* formation of ookinetes and functional maturity of *Plasmodium berghei* gametocytes. *Parasitology* 91, 19–29
- 23 Gygi, S.P. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999
- 24 Nirmalan, N. *et al.* (2004) Quantitative proteomics of the human malaria parasite *Plasmodium falciparum* and its application to studies of development and inhibition. *Mol. Microbiol.* 52, 1187–1199

New compounds outperform existing treatments for both malaria and cancer

Early testing of several new compounds designed by researchers at The Johns Hopkins University in Baltimore (<http://www.jhu.edu>) has revealed that they are a safer and more potent treatment against malaria and some forms of cancer than the current 'gold standard drugs'.

Two of the new compounds outperformed sodium artesunate – the gold standard for malaria treatment – in rodents. Testing of the compounds in rodent models for human prostate cancer produced equally promising results with one compound proving to be three times as powerful as adriamycin – a gold standard anticancer drug.

The compounds in question, named trioxanes, mimic the mechanism of action of artemisinin, the active agent in the *Artemisia annua* plant, which has been used in China as a herbal remedy for malaria and other fevers for thousands of years. The mode of action for drugs such as artemisinin and other antimalarial trioxane drugs is to cause the malaria parasites to self-destruct. As Gary Posner, leader of the research team, explained 'We know that the malaria parasites digest hemoglobin in order to get nutrients, and in the process they release heme. When the heme encounters the peroxide bond, a chemical reaction occurs. Powerful chemical species such as carbon-free radicals and oxidizing agents are produced, harming and eventually killing the parasites'.

The development of new and more effective antimalarial drugs is a worldwide priority. The most commonly fatal strain of the malaria parasite began showing formidable resistance to current treatments decades ago. Coupled with their cancer-fighting properties, the researchers are promoting the compounds as dual-use drug candidates.

For further information, please go to http://www.jhu.edu/news_info/news/home04/aug04/posner.html

Compiled by Anthony Li (A.Li@elsevier.com).