Technologies for the study of gene and protein expression in *Plasmodium*

Daniel J. Carucci

Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, MD 20910, USA
(caruccid@nmrc.navy.mil)

With the imminent completion of the genome sequences of several species of *Plasmodium*, attention is now turning to the exploitation of these genomic sequence data for vaccine, drug and diagnostic development. Several technologies have been developed over the past decade to assist in the determination of gene and protein expression on a global scale. Of these, DNA microarrays, novel high-throughput proteomic technologies and recombinational cloning technologies are lowering the barrier to functional genomic studies in *Plasmodium*. Of equal importance is the capacity to manipulate, store, retrieve and analyse the tremendous quantity of data generated from these genomic studies. This paper will address the use of these technologies as well as some of the computational tools that will be ultimately required to adequately study gene and protein expression in *Plasmodium*.

Keywords: *Plasmodium*; genomics; functional genomics; DNA microarrays; proteomics

1. BACKGROUND

In the 6 years since the first determination of the entire genetic blueprint of a free-living organism (Fleischmann *et al.* 1995), there has been an explosion of data derived from genomic sequencing projects from academic, government and private institutions. At the time of the first draft of this paper, there were 43 published microbial genome sequences and 158 microbial genome projects underway in the public domain (<http://www.tigr.org>) and it is probable, given the current rate, that before this paper is published many more will be initiated. The year 2001 saw the publication of the completed human genome (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001), a feat that some consider will be the most significant scientific accomplishment this century for these data hold the promise of revolutionizing approaches to modern medicine. The genetic basis for many diseases, the discovery of new drugs, the understanding of cancer and the design of new vaccines may all be founded on the data generated from the human genome project and projects like it. Technologies designed to exploit these sequence data have developed even more rapidly. Gene chips, DNA microarrays and proteomics are a few of the terms that are becoming commonplace in typical laboratories throughout the world. These 'functional genomics' technologies are being applied to answer questions regarding the expression of genes and proteins and will certainly address fundamental biological questions.

For those developing methods for controlling the impact of infectious diseases, genome sequencing and the functional genomics are beginning to play a significant role in shaping these control strategies. I will attempt to outline some of the technologies that are already being used or are under development and how these may aid us in understanding pathogen biology, the development of

new vaccines, the design of new drugs and elucidating the mechanisms of drug resistance.

2. MALARIA GENOMICS AND FUNCTIONAL GENOMICS

The first chromosome sequence published from an eukaryotic human pathogen was chromosome 2 from *Plasmodium falciparum* (Gardner *et al.* 1998). Two complete chromosomes and their annotation have now been published and the genome sequences for the remaining 12 chromosomes should be published by mid 2002. From these first publications and from the early released sequence data, it became clear that an enormous amount of information would soon be available that would assist in elucidating parasite biology, uncover mechanisms of host(s) immune-system evasion, determine potential parasite receptors for the invasion of multiple-cell types and identify mechanisms of drug action and resistance. What was not yet clear was how best to utilize these data. Functional genomic studies of bacteria, yeast, *Caenorhabditis elegans*, *Drosophila melanogaster*, mice and humans have developed resulting in elegant methods to determine gene and protein expression on a genome-wide scale and have begun to address fundamental biological questions using these systems.

3. DNA MICROARRAYS FOR STUDYING GENE EXPRESSION

One widely used method for studying transcriptional changes across a genome is DNA-microarray hybridization. DNA microarrays consist of nucleic acid, either amplified DNA or more recently synthesized oligonucleotides, printed onto a derivitized glass surface (typical glass microscope slide) cleaned and coated with either poly-L-lysine or amino silane. A robotic print head

spots a few nanolitres of each specific amplified nucleic acid (gene) from 96 or 384 well plates onto precise locations on the glass slide in a grid pattern. The identity of the gene on the glass slide is determined from its plate well address and its location on the slide. Complementary DNA is synthesized from total or messenger RNA from two different biological samples (i.e. a drug-treated sample and a control) to incorporate one of two fluorescent nucleotides (Cy3- or Cy5-dNTP (deoxynucleotide triphosphate)) into the first strand cDNA synthesis. The two cDNAs (one control and one test) each of which are differentially labelled with the Cy-nucleotides are allowed to hybridize to the surface of the DNA microarray. After washing the slide, a scanning dual-channel confocal laser is used to measure the relative fluorescence intensities at each spot address, giving the relative amounts of each labelled cDNA and therefore the relative amounts of each initial RNA transcript from the two samples. DNA microarrays created using either pre-existing EST or GST (glutathione S-transferase) libraries as templates for amplification, *de novo* gene-specific amplification by PCR or printed synthesized oligonucleotides have their relative advantages. *Plasmodium falciparum* DNA microarrays based on pre-existing libraries (Hayward *et al.* 2000; Mamoun *et al.* 2001), from oligomers (J. Derisi, unpublished data) and from chromosome-specific PCR products (D. Carucci, unpublished data) have been constructed. Since only the blood stages of *P. falciparum* can be cultivated *in vitro*, this stage of parasite development has been the most amenable to study by DNA microarrays. These arrays are being used to study changes in gene expression that occur with *in vitro* cultivation, the effects of drugs on parasite growth *in vitro*, the mechanisms of drug action and resistance and others. In general, 'significant' levels of change are typically reported when one sample is over- or underexpressed by two-fold or more. However, great care must be taken in the interpretation of these data, and additional experiments must be carried out to confirm the biological relevance of these determinations. One tool that is often overlooked in many DNA microarray publications, is the use of traditional statistics. Although multiple-replicate samples are frequently studied, the variance between these data is often not reported. Furthermore, most studies are carried out on slides where each gene is printed only once or at most twice on any single DNA microarray. Where possible, triplicate or quadruplicate samples ought to be included on each array to control for local areas of poor-quality hybridization or poor-quality spotting. Replicate experiments need to be run using replicate biological samples (not replicate runs from the same RNA sample) and if possible more than one region of each gene should be included on the arrays. Finally, rigorous statistics are needed to determine if the difference in fluorescence intensity and therefore gene expression is significant. One method is the use of the standard paired *t*-test on the mean of each triplicate sample to determine if a change in gene expression is statistically significant ($p < 0.05$). Only by applying these types of statistical analyses or others can one begin to determine the relevance of observed changes in gene expression.

4. PROTEIN-EXPRESSION TECHNOLOGIES

Protein-expression studies involving the use of two-dimensional gel electrophoresis is widespread and has formed the foundation for much proteomic research. In this system, separated proteins are usually transferred to a solid support, visually excised and identified by mass spectrometry or microsequencing. Although this method is a mainstay in proteomics, its reliance on few, soluble proteins, proteins of mid to low molecular weight and reduced reproducibility limit its utility to large scale, high-throughput proteomics. One method that overcomes some of these limitations is the combination of liquid chromatography and tandem mass spectroscopy. In this system, proteins are identified from a complex mixture by first cleaving proteins with an endopeptidase (trypsin), separating the resultant peptides on a microliquid chromatography column (100 μm in diameter) and determining the origin of the peptide fragments by comparing their mass spectral 'fingerprints' to the theoretical fingerprints derived from a protein database (Eng *et al.* 1994; Link *et al.* 1999; Washburn & Yates 2000; Yates 2000). The complement of proteins contained within the original protein mixture is determined by the set of multiple peptides from each protein achieving a threshold correlation score. This method, LC-MS-MS, has been recently used to identify over 1400 proteins from the *S. cerevisiae* proteome including over 100 integral membrane proteins (Washburn *et al.* 2001). This method has been used to identify a set of over 1000 proteins from several stages of *P. falciparum*, including mosquito and blood stages of parasite growth (L. Florens, unpublished data). One particularly attractive feature of this approach to protein identification is that membrane and insoluble proteins are not excluded from analysis as is generally the case in traditional two-dimensional protein gels. Insoluble portions of protein extracts are treated with cyanogen bromide, cleaving hydrophilic portions of integral membrane proteins. These solubilized loops of protein are then digested with trypsin and analysed by LC-MS-MS. The resulting identified peptides are then used to identify the presence of the original protein. This technique was recently used to map peptides to yeast integral membrane proteins (Walhout *et al.* 2000).

By targeting protein identification to purified red-blood cell ghost membranes, this method may be used to identify the complement of parasite-derived proteins exported to the surface of the infected red cell. This set of proteins would form an important set of targets for the development of vaccines against the infected erythrocyte. The parasite stage(s) thought to be most important for the development of vaccines to produce sterile protective immunity are the pre-erythrocytic stages, sporozoite and infected liver stage. Injected sporozoites are in the circulation for minutes and thus may only be briefly exposed to clearance by the immune system. Nevertheless, high titre antibodies against this stage of the parasite may be an important factor for limiting hepatocyte invasion, if not killing. However, the infected liver stage is quiescent from a pathological point of view for nearly 6 days of development. In fact, infection with sporozoites from radiation-attenuated infected mosquitoes confers 95% protection from subsequent challenge. Since it is thought that this

protection is primarily due to CD8⁺ T cells directed at the infected hepatocyte, determining the set of parasite proteins expressed at this stage of parasite development provides a critical set of proteins for pre-erythrocytic stage vaccine development. Unfortunately this parasite stage is almost inaccessible for study. Recently, individual infected hepatocytes have been isolated by LCM. Sufficient material can be collected over a few days to produce RNA for RT-PCR studies and protein for Western blotting (J. Sacci, unpublished data). With the exquisite sensitivity of mass spectrometers, particularly the most recent generation machines, it is conceivable that LC-MS-MS can be used to identify large numbers of proteins from this critical stage of the parasite for vaccine development. LC-MS-MS, therefore, provides a high-throughput method for protein identification that may overcome some of the limitations of two-dimensional protein gel electrophoresis.

5. MOLECULAR TOOL DEVELOPMENT

Although genomic sequencing, DNA microarrays and high-throughput proteomics provide attractive technologies for delving into parasite biochemistry and may provide clues for drug and vaccine development, molecular reagents to conduct critical experiments are still required. For many of these studies, DNA plasmids containing whole or partial genes will be the basis of these studies. Protein-expression plasmids, DNA vaccines and transfection vectors will be required for each gene or gene fragment to be studied. Traditional plasmid cloning of each of the estimated 5000 genes from *P. falciparum* is a costly and time-consuming venture. Further it is probable that each of these 5000 genes would, in fact, be required to be cloned into multiple vectors, depending on the experiment in question. Finally, the highly biased genome of *P. falciparum* to adenine and thymine (ca. 80%) is generally unstable in standard plasmid vectors in *Escherichia coli*; requiring the screening of often dozens of potential recombinant clones for each successful clone obtained. A novel approach based on bacteriophage lambda recombination into bacterial chromosomes has been developed that may overcome the technical difficulties of large-scale traditional cloning. This system, Gateway cloning system (Life Technologies, Gaithersburg, MD, USA), has been used to clone large numbers of ORFs from *S. cerevisiae* (Walhout *et al.* 2000). Clone sets from human, yeast and others are being developed in this recombination-based cloning system. Pilot studies have demonstrated that the Gateway system may provide the platform for large-scale cloning of genes from *P. falciparum*; early results indicate that these clones are more stable and more easily obtained than by traditional cloning methods (J. Aguiar, unpublished data). A further advantage of this cloning strategy is that since plasmid inserts can be readily transferred from one plasmid clone to the next at high levels of efficiency (approaching 100%), creating multiple-plasmid constructs (GFP-fusion vectors, His-tagged vectors, transfection vectors, etc) of the same insert is straightforward. Using this method, it is now conceivable to produce multiple-plasmid clones against thousands of identified genes for substantially less cost than by traditional methods.

6. CONCLUSIONS

The completion of the *P. falciparum* genome and the genomes of both human and mosquito hosts offers an unprecedented opportunity to study in detail parasite biochemistry, host-parasite interactions, mechanisms of drug action and resistance, as well as advance vaccine development. Similar to the way that technologies were developed in the mid 1980s to increase the speed, accuracy and breadth of DNA sequencing, as well as to decrease costs, technologies are currently being developed to exploit the enormous quantity of genetic sequence data generated from an ever-increasing number of organisms. Compounding efforts to develop new malaria intervention strategies are the unique problems related to accessing the key stages of parasite development, particularly in the pre-erythrocytic stages. Although establishing partnerships with well-funded functional genomics laboratories in yeast genetics, cancer biology and others will allow early access to newly developed technologies and has already provided the foundation for much of the current state of malaria functional genomics, strategic investments in equipment, personnel and other infrastructure in laboratories studying malaria will be necessary. The advances in functional genomics as a direct result of large-scale genomic sequencing are certain to be pivotal to biomedical research in the 21st century.

The author thanks Dr Laurence Florens, Dr John Yates, Dr Josh LaBaer, Dr Adam Witney, Dr Joao Aguiar and Robert Anthony for their comments and access to preliminary experimental results. The work was funded by the Naval Medical Research Center Work Units 61102AA0101BFX and 611102A0101BCX and the US Army Medical Research and Material Command. The opinions expressed are those of the author and do not reflect the official policy of the Department of the Navy, Department of Defense or the US government.

REFERENCES

- Ben Mamoun, C. (and 11 others) 2001 Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol. Microbiol.* **39**, 26–36.
- Eng, J. K., McCormack, A. L. & Yates, J. R. 1994 An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrometry* **5**, 976–989.
- Fleischmann, R. D. (and 39 others) 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Gardner, M. J. (and 24 others) 1998 Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132.
- Hayward, R. E., Derisi, J. L., Alfaradhi, S., Kaslow, D. C., Brown, P. O. & Rathod, P. K. 2000 Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* **35**, 6–14.
- International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. & Yates III, J. R. 1999 Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17**, 676–682.

- The Institute for Genomic Research (TIGR) (<http://www.tigr.org>)
- Venter, J. C. (and 273 others) 2001 The sequence of the human genome. *Science* **291**, 1304–1351.
- Walhout, A. J. (and 6 others) 2000 GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328**, 575–592.
- Washburn, M. P. & Yates III, J. R. 2000 Analysis of the microbial proteome. *Curr. Opin. Microbiol.* **3**, 292–297.
- Washburn, M. P., Wolters, D. & Yates III, J. R. 2001 Large-scale analysis of the yeast proteome via multidimensional protein identification technology. *Nature Biotechnol.* **19**, 242–247.
- Yates III, J. R. 2000 Mass spectrometry. From genomics to proteomics. *Trends Genet.* **16**, 5–8.