Entering the post-genomic era of malaria research

Paul Horrocks,[1] Sharen Bowman,[2] Susan Kyes,[3] Andrew P. Waters,[4] & Alister Craig[5]

The sequencing of the genome of *Plasmodium falciparum* promises to revolutionize the way in which malaria research will be carried out. Beyond simple gene discovery, the genome sequence will facilitate the comprehensive determination of the parasite's gene expression during its developmental phases, pathology, and in response to environmental variables, such as drug treatment and host genetic background. This article reviews the current status of the *P. falciparum* genome sequencing project and the unique insights it has generated. We also summarize the application of bioinformatics and analytical tools that have been developed for functional genomics. The aim of these activities is the rational, information-based identification of new therapeutic strategies and targets, based on a thorough insight into the biology of *Plasmodium* spp.

Keywords: *Plasmodium falciparum*, genetics; genome, protozoan; base sequence; sequence analysis, DNA.

Bulletin of the World Health Organization, 2000, **78**: 1424–1437.

Voir page 1434 le résumé en français. En la página 1434 figura un resumen en español.

Introduction

The development of genome sequencing technologies over the last five years has resulted in a wealth of sequence information, culminating in the recent announcement of a working draft of the human genome. Pathogen genomes, through their smaller size, have been even more tractable to these methodologies and are now well represented in genome science. Although not always suited as "model" organisms, the importance of pathogens in medicine and agriculture has made the exploitation of the sequence databases a high priority. But what does the production of long strings of As, Cs, Gs and Ts actually mean in terms of the alleviation of the burden of disease, particularly in developing countries? So far, genome sequencing has largely been the province of the developed world, where the resources and science infrastructure have allowed the formation of high-throughput sequencing centres. However, the use of sequencing information need not be restricted in this way, provided that resources for training can be met.

Probably the most important aspect of the post-genomic era (i.e. after the sequencing has been carried out) is analysis of the primary sequence data. This is called bioinformatics and embraces a range of theoretical analyses aimed at converting the DNA sequence into biological information. A major part of this discipline involves the identification of genes through a number of processes, from identifying similarities with previously identified genes from other organisms, to the use of computer-derived models based on existing data. Subsequently come predictions of biological function and molecular shape (structural genomics), both of which have scope for development.

Knowing the whole sequence of a pathogen genome also allows researchers to investigate the behaviour of organisms on a much broader basis than was previously possible. Now, instead of studying the effect of drug treatment or differentiation on one or two genes, it is possible to study variation in all the genes at the same time using global transcriptional analysis. Protein patterns may also be examined, or the organism may be genetically modified (transfection), providing a direct link between these biological effectors and gross phenotype. The technology for these experiments has been developed as a direct consequence of the desire to exploit the genome sequence data.

It is not hard to envisage that the ability to identify and characterize the genetic blueprint of pathogens will help us recognize critical elements in the development and pathogenesis of disease-causing organisms and target our research efforts in the production of new therapies. For *Plasmodium falciparum*, genes and proteins acting at specific stages in the life cycle can be identified, their roles tested by

[1] Post-Doctoral Research Assistant, Molecular Parasitology Group, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, England.

[2] Manager, Malaria Sequencing Project, Pathogen Sequencing Unit, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, England.

[3] Post-Doctoral Research Assistant, Molecular Parasitology Group, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, England.

[4] Reader, Department of Parasitology, Leiden University Medical Centre, Leiden, Netherlands.

[5] Senior Lecturer, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, England (email: agcraig@liv.ac.uk). Correspondence should be addressed to this author.

Ref. No. **00-0825**

genetic modification and promising candidates used in vaccine production. Parasite metabolic pathways not present in the host could also be targeted with potent inhibitors that are non-toxic to humans; and parasite drug-resistance mechanisms could also be targeted, giving existing drugs a longer effective lifetime. The sequence of the *P. falciparum* genome will provide many opportunities for research into malaria, but this is only a beginning, with the challenge being to turn those opportunities into effective treatments in the field.

The *Plasmodium falciparum* genome

The genome of *P. falciparum* consists of three discrete components: a linear repeat of a 6 kb element located within mitochondria; a 35 kb circle within a plastid-like structure (the apicoplast); and 25–30 Mb of nuclear DNA (genomic DNA). The nuclear DNA is organized into 14 chromosomes, between 0.75–3.5 Mb in size, as determined by pulse-field gel electrophoresis (PFGE) and electron microscopic counts of kinetochore structures. Indirect evidence for the number of chromosomes also comes from genetic studies showing there are 14 linkage groups.

The nuclear genome is organized in a manner typical of eukaryotes, with linear chromosomes being bounded at either end with telomeric sequences. Genome plasticity, seen in many parasite isolates and identified by size polymorphisms on PFGE, is thought to result frequently from deletions and insertions of DNA within subtelomeric sequences, a region shown to contain ordered repetitive sequence elements.

The *Plasmodium falciparum* Genome Project Consortium

Prior to the advent of yeast artificial chromosome (YAC) technology, there had been relatively little access to the parasite's genome, as its extreme [A+T] content rendered inserts unstable in conventional bacterial plasmid clones. The estimated 80% [A+T]-rich genome could, however, be stably maintained within the pYAC4 construct, as demonstrated by the construction of a number of YAC libraries for different *P. falciparum* clones (*1*). In 1993 a consortium of laboratories distributed throughout the world established the Wellcome Trust Malaria Genome Mapping Project, with the aim of assembling YAC contiguous sequences (contigs) across each chromosome, as well as developing YAC, expressed sequence tag (EST), bioinformatic and genetic mapping technology (*2*). This consortium realized that sequencing of the entire nuclear DNA was a real possibility, yet considered the endeavour fraught with difficulties due to the extreme bias in base content.

Genome sequencing has proved to be a powerful and efficient approach in accessing the complete gene complement for organisms as diverse as *Mycobacterium tuberculosis* (*3*), *Saccharomyces cerevisiae* (*4*) and *Caenorhabditis elegans* (*5*). The advantages offered by such a tool in the investigation of human malaria eventually resulted in pilot projects being established in 1996 at three high-throughput genome centres, to establish whether sequencing the entire genome was possible: the Sanger Centre (England); The Institute for Genomic Research (TIGR) Malaria Program, Naval Medical Research Centre (NMRC) (USA); and Stanford University (USA). The Wellcome Trust, the Burroughs Wellcome Fund, the National Institute of Allergy and Infectious Diseases and the US Department of Defence provided funding for the pilot projects, and following their success these agencies agreed to fund the entire sequencing effort (*6*; see Table 1 for progress). Associated with this work are a number of other groups supporting the efforts of the high-throughput centres in a range of activities, including generation of chromosomal material; additional mapping information; testing bacterial strains more tolerant of [A+T]-rich DNA; and the provision of a repository for *P. falciparum* reagents (MR4, http://www.malaria.mr4.org/).

A similar strategy is being used by all of the high-throughput sequencing centres, with individual chromosomes being excised from pulse-field gels, cloned as small inserts into a double-stranded vector, and sequenced in the forward and reverse directions to generate read-pair information which is used in gap filling. Sequence-tagged site and simple sequence-length polymorphism microsatellite markers from the HB3xDd2 genetic cross (*7*), together with the optical map (*8*) of ordered restriction fragments, are used to position contigs on each chromosome, or to confirm that sequence data has been assembled correctly. In addition, to help assign and order sequences originating from a particular chromosome region, the groups at the Sanger Centre and Stanford University use a shotgun skim (1–2 fold coverage) of YAC clones selected from the chromosomal YAC contigs, generated by the original *P. falciparum* mapping project. All three sequencing centres aim for an error rate of less than 1 base in every 10 000 bases.

Once a section of chromosome sequence is finished it is analysed to identify a number of features, including putative protein-coding regions, tRNAs and repetitive sequences. Database searches are performed to identify similarities to protein and EST sequences. Further analyses are performed to determine whether the predictions have protein domains, signal sequences, putative membrane-spanning regions or any other distinctive features. A number of computing tools such as Hexamer/Genefinder (R. Durbin, P. Green, L. Hillier, unpublished software, 1998)), and GlimmerM (*9*) are available to assist in the identification of protein-coding regions, but many other gene prediction programs exist or are currently being developed.

These tools are efficient at identifying single, large, open reading frames (ORFs) or genes with two or three relatively large exons. However, the current

Table 1. **Progress summary for the Malaria Genome Project**

Chromosome	Size (Mb)	Status[a]
14[b]	3.4	Late Closure
13[c]	3.2	Closure
12[d]	2.4	Late Closure
11[b]	2.4	Closure
10[b]	2.1	Closure
9[c]	1.8	Shotgun in progress
8[c]	1.7	Shotgun in progress
7[c]	1.7	Shotgun in progress
6[c]	1.6	Shotgun complete
5[c]	1.4	Closure
4[c]	1.2	Late Closure
3[c]	1.06	Finished
2[b]	0.95	Finished
1[c]	0.7	Late Closure

[a] Shotgun = random phase of sequencing where no selection of clones is used.

Closure = the shotgun phase leaves gaps in the chromosomal sequence and these have to be filled using combinatorial approaches based on the sequence information flanking the gaps. Sometimes the sequence data do exist but have to be extracted manually from the computer database.

Late Closure = sometimes a small number of gaps remain which are refractory to high-throughput sequencing methodologies and these need highly focussed approaches to fill them.

[b] The Institute for Genomic Research, Malaria Program, Naval Medical Research Centre.

[c] The Sanger Centre.

[d] Stanford University.

generation of gene prediction software produces conflicting data when predicting multi-exon genes, particularly those with small exons (*10, 11*). These conflicting gene predictions are currently being tested experimentally by reverse-transcription polymerase chain reaction (RT–PCR) assays (Fig. 1). Annotation of current and future *P. falciparum* sequences is therefore an ongoing process, with predictions and annotations being refined as more information, such as the RT–PCR and EST sequencing data, becomes available.

Analysis of the *Plasmodium falciparum* genome

A total of 424 predicted protein-coding genes and 3 tRNA genes have been identified on chromosomes 2 and 3, the two chromosomes for which sequencing has been completed (*12*, *13*). Approximately 37% (158 genes) of these genes have a readily identifiable homologue in another species. Such similarity allows a function to be implied, with many of those identified being involved in parasite metabolism. Interestingly, a comparison of the predicted protein sequences with their homologues in other species showed that the majority of *P. falciparum* proteins have insertions of low complexity sequences, often runs of a single amino acid residue (typically asparagine, lysine or glutamic acid), or tandem arrays of a short peptide repeat sequence. Examples of such regions have been identified previously, and are polymorphic between different parasite isolates.

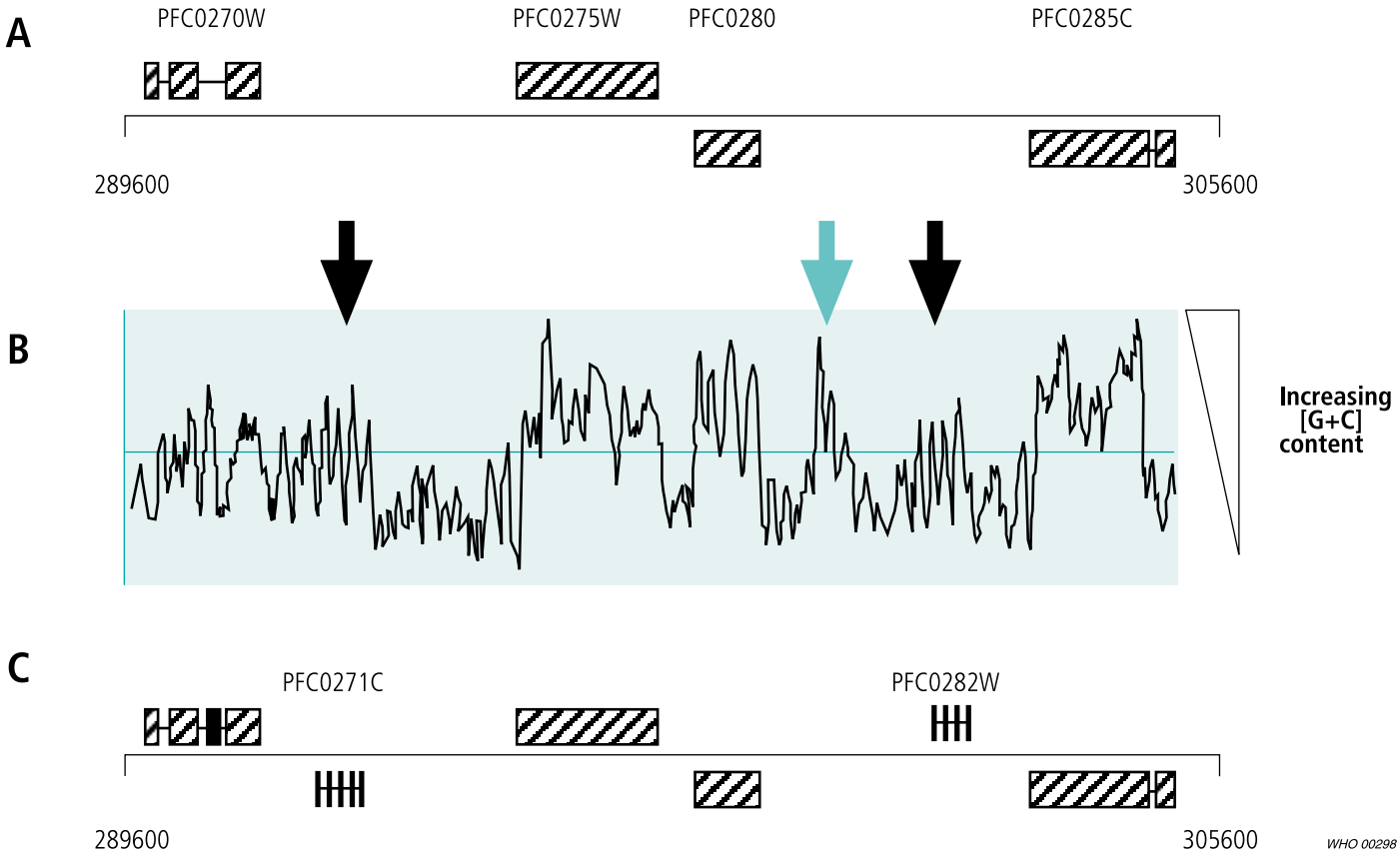
P. falciparum contains two organelles thought to have arisen through endosymbiotic events: mitochondria (*14*) and apicoplasts (*15*, *16*). The unique nature of the apicoplast and the essential functions that it carries out make it a prime candidate for antimalarial drug development. Like many extranuclear elements, many of the genes for the organelle function have become nuclear encoded. Examination of the predicted proteins encoded on chromosomes 2 and 3 identified several that have a putative apicoplast signal sequence. This indicates that the apicoplast contains type II fatty acid synthase systems, typically associated with bacterial and plant plastids (*17*, *18*). This supports the hypothesis that the apicoplast is algal in origin and provides a very specific target for rational drug design.

Other proteins likely to play a critical role in parasite metabolism are also easily identified from searches of the *P. falciparum* databases. Using this approach, Jomaa et al. (*19*) identified two *P. falciparum* proteins with similarity to enzymes involved in the 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid synthesis, which is used in green algae and some bacteria, but not in animals. Antibiotics developed to inhibit this pathway have already been shown to have antimalarial activity in a rodent model system and in *P. falciparum* culture in vitro.

Database searches permit the "virtual" identification of *P. falciparum* homologues to proteins from other species, but comparative analysis of sequence organization (*12*, *13*) has played a significant role in revealing subtelomeric coding sequences that would otherwise have gone unnoticed. A closer examination of the four available *P. falciparum* subtelomeric sequences shows that the order of both repetitive sequences and the variant multigene family members are conserved (Fig. 2). Members of the *var*, *rif, stevor* and *Pf60* multigene families are present at all four telomeres, and new telomere-associated multigene families have also been identified (conserved telomere-encoded proteins (CTPs)). While *var*, *Pf60* and *stevor* had already been relatively well-characterized, *rif* required a more detailed analysis.

Var comprises a highly polymorphic gene family of roughly 50 copies per haploid genome (*20–22*). It codes for PfEMP-1 (*P. falciparum* erythrocyte membrane protein-1) variants expressed on the surface of red blood cells, from the late ring stages of infection through schizogony. Each *var* gene is composed of two exons (Fig. 3). Exon 1 codes for the highly variable extracellular portion of PfEMP-1, including between one and seven Duffy-binding-like (DBL) domains, with at least one cysteine-rich interdomain region (CIDR) (*23*). The 3′ end of exon 1 codes for the semiconserved transmembrane region, and exon 2 codes for the semiconserved cytoplasmic region of the protein. PfEMP-1 is involved in cytoadherence of infected red cells to a range of different host receptors on endothelial cells and in binding to uninfected red cells, forming rosettes. The cytoadherent or rosetting phenotype of an individual parasite-infected red cell

Fig. 1. **Testing gene models in *Plasmodium falciparum* – annotation of the *Plasmodium falciparum* genome is a continuing process.** A) The region on chromosome 3 between bases 289 600 and 305 600 initially shown to contain 4 ORFs (PFC0270W to PFC0285C, hatched boxes). B) However, analysis of the [G+C] content of this region in overlapping 100 bp segments, using the Artemis viewer, indicated three further regions of higher-than-average [G+C] content (black and green arrows). C) Subsequent RT-PCR analysis of asexual parasites identified two further genes (PFC0271C and PFC0282W, filled boxes), as well as a modification to the PFC0270W gene model. Further analysis using other developmental stages may identify additional genes, particularly associated with the [G+C] peak shown by the green arrow in B).



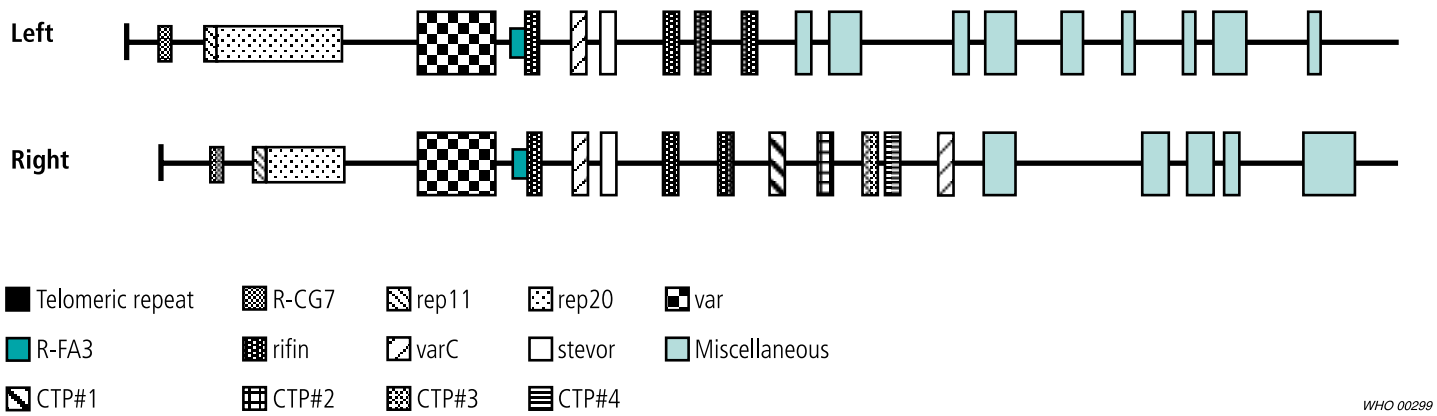
has been correlated with severity of disease and disease pathology (reviewed in *24*).

The *Pf60* family members contain a 3′ exon highly similar to the *var* exon 2 sequence (*25*, *26*), and have multiple possible 5′ exon organizations. One *Pf60* protein, 6.1, is expressed in the nucleus of late asexual blood stage parasites, is composed of 7 exons, and employs a mechanism for reading through an internal stop codon (*27*). Hybridization-based analysis of *Pf60* suggested there were 140 copies per haploid genome, but as the sequences are highly similar to *var*, an adjusted estimate would be roughly 90 *Pf60* genes per haploid genome.

Rif was originally described as an ~1-kb DNA fragment repeated in the *P. falciparum* genome, an anomalous open reading frame (ORF) with no logical methionine start site (*28*). There was little point in hunting for upstream exons, since the dogma at that time was that most *P. falciparum* genes did not have introns. The sequence was eventually spotted between two *var* genes (*22*), and from early releases of genome sequences *rif* was recognized as a relative of *stevor* (*29*). This analysis identified the small upstream exon for both *stevor* and *rif* sequences (Fig. 3), and showed that both sequence types contain predicted transmembrane regions, which would orient the predicted protein as a loop on the outer surface of a cell membrane. The *rif* sequences are distinct from, and much more polymorphic than, *stevor* with an estimated 200 copies of *rif* per haploid genome (*13*, *30*), and roughly thirty-four members of the *stevor* family (*29*).

To date, it is uncertain where and when the STEVOR proteins are expressed (*31*), although the relative lack of polymorphism suggests that they are not likely to be expressed on the surface of red blood cell. However, for *rif* the large number of highly polymorphic copies suggested a location exposed to host immune/selective pressures. Indeed, it was shown that *rif* sequences code for clonally variant 35–44 kD RIFIN proteins expressed on the surface of infected red blood cells (*30*), and that antibodies to the proteins are detectable in sera from immune individuals (*32*). Although RIFINs and PfEMP-1 share the same cellular localization, RIFIN function remains unclear. While *var* genes are transcribed during all ring stages of development, and *rif* genes are only expressed for a short time at the late-ring/early pigmented

Fig. 2. **Subtelomeric organization of gene families in *Plasmodium falciparum*.** The telomeres of chromosome 3 are shown, with sequence repeats and multigene family members indicated as shaded boxes. A comparison of chromosome ends demonstrates that the order of tandem repeat sequences is conserved, although copy number can vary. Some conservation in the order and orientation of multigene family members can also be seen. For example, the CTP genes (conserved telomere-encoded protein) are highly similar to a set of genes seen in the right subtelomeric region of chromosome 2.



trophozoite stage, both proteins are detected on the red cell surface at roughly the same time (early trophozoites) (*25*, *33*). Further studies are necessary to determine whether these proteins are functionally linked as well as physically linked within the genome.

The final gap to be sequenced on chromosome 3 covered a region of extreme [A+T] composition: 97.3% for 2.6kb. Subsequent comparison with the sequence of chromosome 2 identified a region similar in both composition and length. Both [A+T]-rich regions occur in gene-sparse areas of the chromosomes, forming part of the longest intergenic region on chromosome 3. Closer analysis revealed the presence of chromosome-specific repetitive sequences. Thus, their structure and extreme [A+T] composition suggest these regions are candidate centromeres. [A+T]-rich central cores are present in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* centromeres and the latter contain complex, chromosome-specific sequences. Although, to date, these regions have not been demonstrated to function as *P. falciparum* centromeres, a third example of this structure has recently been identified on chromosome 1.

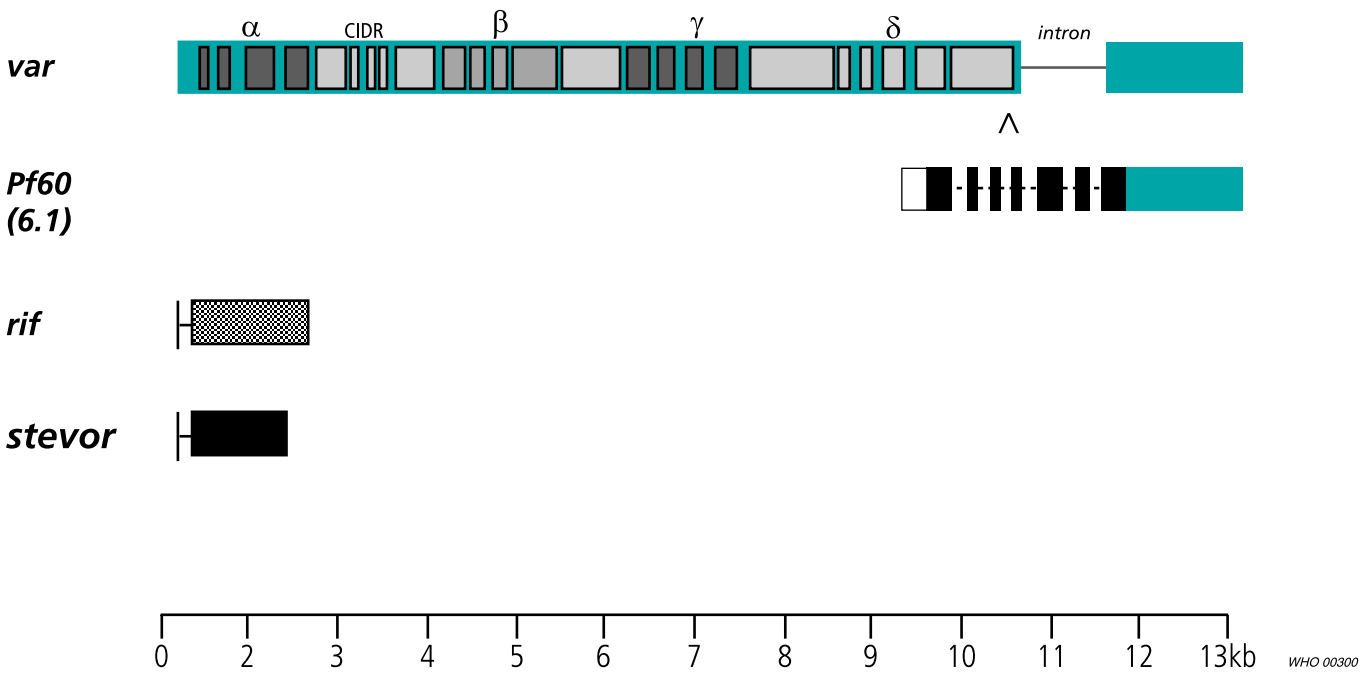
Functional genomics

The availability of an increasing number of pathogen genome sequences has strengthened the impetus for "global" investigation. The paradigm for these studies has been set by yeast researchers, who have had access to the full genome sequence of *S. cerevisiae* since 1996 (*4*). Although implicit in genomic sequencing, the development of techniques to study functional parameters across whole genomes has evolved into a new field of research termed "functional genomics". For *P. falciparum*, a number of technical difficulties remain, but researchers are already applying functional genomics to the biology of this parasite.

Bioinformatics

Bioinformatics covers both sequencing and functional genomics in converting raw genome sequence data into useful information for the biologist. Probably the most important aspects are the search engines that can screen vast amounts of information for significant similarities between sequences, and the algorithms used to predict protein-coding regions (see above). As far as the search engines are concerned, several sites exist on the web that will allow the user to search for homology to a test sequence (see Table 2), returning the results as multiple sets of alignments. Although there are some errors in defining genes "in silico" (i.e, a gene model predicted by computer), it is clear that these techniques have generated an enormous number of potential gene functions that can be tested in the laboratory. Perhaps the easiest forms of this are the annotations of complete chromosome sequences that are based on high-quality sequence information. The resulting tables of genes can be divided into those that have confirmed function in *P. falciparum*; sequences with homology to known genes in other organisms; sequences with homology to genes of unknown function in other organisms; and sequences specific to *P. falciparum* of unknown function. One of the underlying assumptions is that the annotated sequence contains all the genes and that the predictions for the coding sequence are correct. While this is largely true, there has been some discussion about "missing" genes and the incorrect identification of splice sites. Therefore, an essential part of any bioinformatic system is the provision of software for browsing genomic sequence data so that

Fig. 3. **Exon/intron structure of subtelomeric genes – schematic showing exon–intron structure for the two superfamilies *var/Pf60 and rif/stevor.*** This example of *var* exon1 contains four DBL-domains (indicated α, β, γ, δ) and one CIDR region. Regions of semiconserved sequence are indicated with dark grey bars, otherwise the sequence is highly polymorphic. The transmembrane region is indicated (^). *Pf60* type 6.1 has 7 exons, the seventh being highly similar to *var* exon2; *rif* and *stevor* are each composed of two exons, and are of similar size.



independent laboratories can make their own judgements on coding sequences.

Access to information is possible through the release of sequence data prior to publication, which has already made a significant contribution to malaria research. However, different web sites contain different parts of the sequence information from different chromosomes. The site at the National Center for Biotechnology Information (NCBI, see Table 2) has attempted to collate all the available information for *P. falciparum* (as well as that for other *Plasmodium* species and the related apicomplexan parasite *Toxoplasma gondii*). A recent development, funded by the Burroughs Wellcome Fund, has established a full database (http://www.PlasmoDB.org). In its current form, PlasmoDB provides: views of finished and annotated sequence in both web-based and CD format; a relational database that should facilitate entry and analysis of expression data; a "BLASTable" database containing the most recent genomic sequence information (finished and unfinished); text-queriable results from a complete BLAST search of all *P. falciparum* data against the GenBank and EMBL databases; and various other data-mining tools. In addition to the obvious interest of PlasmoDB to the malaria research community, the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases has made a commitment to provide on-line "help-desk" support for PlasmoDB (and other parasite databases), through the training of scientists/bioinformaticians from developing countries.

Microarrays

Microarrays are high-density arrays of DNA targets on glass or filter supports. These have been used in several studies, most notably for *S. cerevisiae*. There are essentially two formats for the DNA targets: DNA fragments (usually generated by PCR) or oligonucleotides. PCR-derived targets for *P. falciparum* have been produced from genome sequence data (D. Carrucci, personal communication, 1998), or by the amplification of sequences from a mung bean nuclease library. In the mung bean nuclease library study (*34*), arrays were made from a random library and screened with

Table 2. ***Plasmodium falciparum* bioinformatic web-site URLs**

http://www.sanger.ac.uk/Projects/P_falciparum (Sequence data)
http://www.tigr.org/tdb/edb/pfdb/pfdb.html (Sequence data)
http://sequence-www.stanford.edu/group/malaria/index.html (Sequence data)
http://www.PlasmoDB.org ('Official' database of the Plasmodium Genome Consortium)
http://www.ncbi.nlm.nih.gov/Malaria (Genetics/bibliography/sequence data)
http://www.ebi.ac.uk/parasites/parasite-genome.html (General site/proteomics)
http://sites.huji.ac.il/malaria/(Metabolic pathways)

mRNA from asexual (trophozoite) and sexual (gametocyte) stages. The RNAs were labelled with different fluorochromes and allowed to hybridize to the arrays. Genes transcribed during asexual stages were labelled green, those transcribed during sexual stages red, and those transcribed during both stages yellow (green + red). From this single experiment the authors identified several developmentally regulated genes, including some which had been previously identified, confirming the efficacy of this approach.

The potential for these types of experiments is considerable, but will require careful data handling and analysis. Hayward et al. (*34*) identified their candidate genes by sequencing the inserts after hybridization, but once the genome has been completely sequenced the address of each ORF will be stored as part of the array information. An array covering all of the genes in the *P. falciparum* genome in duplicate would have approximately 12 000–14 000 spots; therefore every experiment would generate at least 14 000 quantitative data points per probe, in addition to the array baseline information. It is not hard to see that without a highly efficient database, the efficiency and accuracy of any analysis would be greatly reduced, particularly since transcriptional changes are likely to be defined by "clusters" of genes (*35*), which makes the contribution of individual genes difficult to interpret. The database must be accessible by the research community to take full advantage of the results and, as seen for *C. elegans* (http://www.wormbase.org), a centralized database is essential.

Transcriptional analysis for *P. falciparum* is also being attempted by Serial Analysis of Gene Expression (SAGE, *36*; D. Wirth, personal communication, 1998)). This technique uses PCR amplification of mRNA to derive short sequence tags that are unique to each gene. The tags are concatenated into long strings and cloned into suitable vectors. By sequencing many tags, it is possible not only to identify what gene is being expressed, but also to quantify the level of expression by recording the number of times specific sequence tags are present. The handling is more complicated than hybridization to arrays, but the major advantage of this technique is that relatively small quantities of material can be used, which is particularly important where samples may be limited (e.g. field samples). For both techniques it is important that technology transfer is carried out, including not only the mechanics such as robotics and image analysis, but also an appreciation of the assumptions and limitations. For example, neither technique can detect the products of alternative mRNA splicing events.

High-density arrays are not exclusively for use in transcriptional analysis, but also have applications in genome analysis, particularly as oligonucleotide arrays. The high degree of specificity of oligonucleotide hybridization allows a high level of discrimination, including single nucleotide substitutions. However, their use is precluded prior to the completion of the *P. falciparum* sequencing project due to the cost of the chemical synthesis. One example of this type of approach comes again from *S. cerevisiae* (*37*), in which allelic variation throughout the genomes of two strains of yeast was identified solely by hybridization of their genomic DNA to oligonucleotide arrays. The application of this technology to natural populations of parasites will be a powerful tool for molecular epidemiology. However, information from many *P. falciparum* genomes will need to be collated to fully cover the true extent of highly variant and recombinogenic genes such as *var*.

Proteomics and structural genomics

One of the greatest technical challenges in functional genomics concerns the analysis of protein expression (proteomics). Although genome-wide transcriptional analysis has produced very useful information, it is clear that the correlation between mRNA and protein levels is not perfect (*38*). Techniques for identifying proteins in general have developed rapidly over the last few years. However, the major advance with regard to genome sequencing has come through the use of mass spectrometry for protein analysis and characterization, by combining the mass (and therefore amino acid composition) of peptide fragments with a database of all the available peptide combinations derived from the genome sequence data (see *39* for review). Using these techniques it is possible to identify individual proteins from complex mixtures resolved by two-dimensional (2-D) electrophoresis or chromatography. This, allied with improvements in the reproducibility of 2-D electrophoresis, has facilitated the differential analysis of protein composition from two populations of cells.

Applications of this technology, for example in the analysis of the effect of drug treatments, have an important place in malaria research. Some practical difficulties also remain, particularly for membrane proteins, which are poorly represented using classical proteomic techniques (*40*). However, technology continues to improve (*41*) and the reward will be a clear definition of the spectrum of proteins involved in critical biological processes in the parasite.

Protein–protein interactions play an important role in cell biology. The tools to investigate this phenomenon, again, have been developed with yeast through the use of the "two-hybrid" system. In this, ORFs are fused to the *Gal4* transcription-activation domain in one yeast strain and screened by mating with a second strain in which specific coding sequences have been fused with the *Gal4* DNA-binding domain. Positive protein interactions are scored by the ability of the yeast progeny to grow on selective media. This procedure has recently been expanded to enable researchers to carry out a comprehensive analysis of the protein–protein interactions of approximately 6000 yeast ORFs (*42*), generating novel information as well as confirming previous specific screens. The extension of this

technique to pathogen genome research will produce a further layer of functional information.

One of the goals of computational biology is to predict the biological function of a protein from primary sequence information. Currently, the clearest manifestation of this is in structural genomics in which the three-dimensional structure of proteins is studied in the context of the amino acid sequence. While much still needs to be done, the increasing number of resolved crystal structures for biological molecules has already resulted in the generation of rules/motifs that can be applied when searching for structural information (*43*). As more information becomes available, including the integration of functional data in terms of active residues, these computer algorithms will improve.

Metabolomics and vaccinomics

One of the many outcomes from functional genomics has been the production of a new vocabulary to cover the many applications of metabolomics and vaccinomics. Thus, the entire mRNA and protein complement of an organism have been termed the transcriptome and the proteome, respectively. In a similar vein, the use of genomic information to facilitate studies of metabolic processes has been termed metabolomics (*44, 45*). Clearly, the use of sequence similarities to identify components of known pathways will have a huge impact on research in this area, but functional screens have also been proposed. For example, by expressing all the predicted ORFs from a genome as heterologous fusion proteins it might be possible to screen for specific enzyme functions. This resource could also be used to determine host protective immune responses through immunization with protein pools. Related studies have been called vaccinomics, in which DNA vaccine plasmids containing ORFs (derived from the sequencing project) are used to immunize experimental animals and, in the case of pathogens, to screen for protection (*46*). Not only could this lead to the identification of vaccine candidates, but the resulting sera could also be used to determine the cellular location of each protein.

As the number of sequenced genomes increases and researchers become more accustomed to thinking in "global" terms, the amount of information generated by functional genomics will expand rapidly. The challenges will then be to develop systems to analyse this information and to maintain a focus on the biological issues. Clearly, for malaria this would be well served through studies on the parasites themselves, facilitated by genetic crosses (*47*) and the technology of genetic manipulation (see below).

Transfection of *Plasmodium falciparum*

Transfection, or the introduction of exogenous DNA into the organism in vivo, potentially provides one of the most powerful tools for the analysis of the parasite's genome, allowing us to specifically address questions relating to gene function. However, prior to the initial demonstration of luciferase expression in the chicken malaria model, *P. gallinaceum*, the ability to modify or disrupt components of the *P. falciparum* genome had eluded researchers (*48*). Following this report, transfection of *P. falciparum* was successfully carried out when a plasmid bearing the reporter gene encoding chloramphenicol acetyltransferase (CAT) was introduced into the readily cultured, intra-erythrocytic stages (*49*). Stable transfection of a plasmid bearing a pyrimethamine selectable marker (*mdhfr-ts*, a mutant dihydrofolate reductase–thymidine synthetase gene), and its subsequent integration via homologous recombination into the genome, soon followed (*50*). The plasmids described in these initial reports, as well as a transgene expression system (expression of a reporter gene from a plasmid stably maintained by virtue of *mdhfr-ts* (*51*)) have provided the basis for all the subsequent studies in *P. falciparum*.

Using both CAT and luciferase reporter genes, a series of studies have addressed the structure and function of *P. falciparum* transcriptional units. *P. falciparum* promoters were shown to conform to a classical bipartite structure: a basal promoter regulated by upstream regulatory factors, with transcript stability directed by 3′ regulatory sequences (*52–57*). Also, the nuclear context in which promoters are placed play a key role in their function, suggesting that phenomena such as stage-specific gene expression, and switching of expression between members of the *var* multigene family, may rely on epigenetic factors, such as chromatin assembly (*58, 59*).

More advanced work, utilizing gene disruption and allelic replacement, has formed the basis of a series of recent reports that examined aspects of *P. falciparum* biology, such as cytoadhesion, gametogenesis and drug resistance (Table 3). Disruption of the gene encoding the knob-associated histidine-rich protein (KAHRP), located within electron-dense knob structures on the surface of infected erythrocytes, showed that this protein is important in their formation (*51*). Follow-up studies demonstrated that the knob structure is critical in supporting PfEMP-1 protein during its interaction with host receptors under fluid flow conditions. Pfg27, a protein expressed early following a parasite's commitment to sexual differentiation, is also essential since disruption of this gene resulted in the abortion of transfectants early in gametogenesis, and resulted in highly vacuolated and morphologically disrupted parasites (*60*).

The roles of mutations within genes conferring resistance to a wide range of antimalarial drugs have also been investigated by altering ORFs rather than disrupting them (*61*). And the roles of dihydropteroate synthase in sulfadoxine resistance, and P-glycoprotein homologue 1 (Pgh-1) in multidrug resistance, were elegantly demonstrated by introducing mutations associated with drug resistance in epidemiological studies into parasites bearing a drug-sensitive back-

Table 3. **Genes modified or "knocked out" in *Plasmodium***

***Plasmodium* gene targeted**[a]	**Phenotype**	**Reference**
Pb CS	Inhibition of sporozoite formation, loss of infectivity	*66*
Pb TRAP	Sporozoites fail to glide and show reduced infectivity	*67*
Pb, Pf CTRP	Ookinetes non-motile and fail to invade midgut epithelia and develop into oocysts	*83–85*
Pf KAHRP	Infected-red blood cells "knobless" and have reduced binding to CD36 under flow conditions	*51*
PF G27	Total (3′ disruption) or significant (5′ disruption) inhibition of gametocytogenesis	*60*
***Plasmodium* gene modified**	**Manipulation and phenotype**	**Reference**
Pf DHPS mutation	Introduction of field-observed mutations give rise to sulfadoxine resistance when introduced	*61*
Pf *pgh1* replacement	Introduction of field observed mutations give rise to mefloquine, halofantrine and quinine resistance. Also resistance to chloroquine in a strain-specific manner.	*62*
Pb TRAP	Replacement with Pf TRAP. Sporozoites can glide, invade salivary glands and are infectious	*86*
Pb TRAP	Replacement with Pf TRAP mutants. TRM^- sporozoites do not glide or invade salivary glands, but remain infectious. A_{mut} sporozoites do not invade salivary glands, but remain infectious	*86*
Pb TRAP	Replacement with Pb TRAP mutant. Cyt △S and △L sporozoites do not glide normally, cannot invade salivary glands and are not infectious. Cyt/MIC-2 sporozoites can glide, invade salivary glands and are infectious	*87*
Pf Acyl Carrier Protein	Green fluorescent protein tag demonstrates need for bipartite leader sequence to correctly target protein to apicoplast and conservation of leader within apicoplasts.	*80*
Pf MSP1	Replaced C terminal with that of *P. chabaudi* MSP-1 demonstrates conservation of function and potential for immune variation.	*88*

[a] Except where indicated, the targeted loci have been disrupted after integration. Genes targeted in *P. berghei* are indicated by Pb or PB and genes targeted in *P. falciparum* are indicated by Pf or PF.

ground (*62*). Moreover, these studies have also been used to demonstrate the reverse situation, where mutations of *cg2*, a candidate chloroquine resistance gene, did not confer resistance in a sensitive parasite background, prompting investigations that identified a more promising candidate (D. Fiddock, T. Wellems, personal communication, 1999).

The primary limitation with this type of experimental strategy is that the intraerythrocytic stage parasites investigated are haploid. Targeted disruption of essential genes inevitably result in the death of the parasite; moreover, any effect on parasite viability will place that proportion of the population at a growth disadvantage. This may be partially overcome by selection, or by the presence of alternative pathways that can overcome the growth disadvantage.

Proposals for a consortium of laboratories, similar to the EUROFAN network for *S. cerevisiae*, to systematically "knock-out" every gene identified by the genome project have considered a wide variety of issues, such as poor parasite viability, low transfection efficiency, the numbers of laboratories able to create mutants, and the unit cost per gene knock-out (*63*). Although there are still many issues to be resolved, it has been agreed that some form of systematic analysis of the parasite's gene complement should take place. However, whether each laboratory would be allocated a share of the genome, or whether a thematic approach is used (based on a laboratory's interest in a particular subject, such as erythrocyte invasion, cytoadhesion or gametocytogenesis) remains to be determined.

Malaria model systems

The lack of in vitro culture systems for most *P. falciparum* developmental stages and the ethical considerations necessary for the use of New World monkeys and chimpanzees highlight the advantages offered by model malaria parasite systems. Transfection of animal malaria models has opened up the entire parasite's life cycle, including those in the invertebrate mosquito host, for critical analysis in a manner not possible with human parasites (*64*). Moreover, the efficiency of the transfection process, facilitating double-cross over gene replacements, and the rapid selection of transfectants in vivo, have resulted in the rapid accumulation of new insights into *Plasmodium* spp. biology.

Following the first reports of transfection in the rodent malarial model *P. berghei* in 1995 (*65*), disruptions of the genes encoding circumsporozoite protein (CS) (*66*) and thrombospondin-related anonymous protein (TRAP) (*67*) demonstrated the value of this approach in the investigation of function: apart from their roles in host cell recognition and invasion, new roles in sporozoite formation and gliding motility were attributed to CS and TRAP, respectively. Recurrent biological themes have

already been demonstrated. For example, CTRP, a TRAP homologue, has been shown to have a role in ookinete motility and thus in the ability to infect mosquitoes.

The readily available and immunologically well-characterized animal hosts allow a full range of host parasite interactions to be investigated. Following the success of this work, transfection of two primate malaria models has been established at the Biochemical Primate Research Centre in the Netherlands. Both *P. knowlesi* (*68*) and *P. cynomolgi* (*69*) have been transfected with entirely heterologous plasmid constructs bearing a *Toxoplasma gondii* DHFR-TS drug selectable marker controlled by *P. berghei* regulatory sequences. One particular advantage of primate malaria models is that they allow investigation of transfected parasites within both their natural and artificial hosts. Within strict ethical limitations, the use of these systems facilitates meaningful analyses of host–parasite interactions and will allow the mechanisms governing the host immune response to be examined, as well as providing an experimental model for evaluating vaccines.

Animal models share many features with their human malarial counterparts, such as life cycles, host-cell restrictions and immune responses. For example, *P. cynomolgi* infection of macaques is a particularly good model for *P. vivax* infection, sharing preference for reticulocytes and hypnozoite formation (*69*). Although we have a substantial pool of knowledge for animal malaria models, particularly with respect to homologous vaccine antigens, we understand comparatively little about these systems at the molecular level. Software for sequencing EST libraries, as well as low-coverage genomic shotgun sequencing, have already provided a fundamental boost to the application of animal models to human malaria biology and should be supported with larger-scale sequencing activities. Use of animal models would allow intensive programmatic approaches to gene function to be considered through the use of knock-out/tagging methods.

Where species of *Plasmodium*, such as *P. knowlesi* and *P. cynomolgi*, provide models for malaria infection and immunity, the closely related apicomplexan *T. gondii* provides a model for many aspects of *Plasmodium* biology (*70*). Transfection systems are well developed in this parasite, with high-efficiency transformation and a number of markers for both positive and negative selection (*71*). This experimentally tractable system has been instrumental in tackling issues such as plastid structure and function (*72*), host cell invasion, drug resistance, virulence and gene expression. Unlike *Plasmodium* spp., which exclusively integrate DNA molecules through homologous recombination, *T. gondii* also permits non-homologous integration, allowing the entire genome to be tagged for a phenotype-based approach to gene function. For this, *Plasmodium* will require a transposon-based system of simple recognition site specificity, such as the *Drosophila mariner* element, which efficiently transfers into a number of eukaryotic genomes including *Leishmania major* (*73*).

Both low- and high-technology developments are needed to continue a systematic assault on the *Plasmodium* genome, such as improved culture methods for certain human and model bloodstage parasites, gamete development and efficient sporozoite culture. These, together with advances in transfection technology, such as new markers, methods to increase transfection efficiencies and adoption of the Tet-repressor system, will combine to make transfection a more reliable and sensitive tool. However, only the coupling of these advances with the knowledge of the parasite's entire gene complement will allow the advances to be fully exploited.

Comparative genomics

Mention has already been made of the difficulties that can be encountered when annotating a complete genome, particularly assigning function to unknown genes and correctly identifying genes that contain numerous exons. Ultimately, the availability of a fully or extensively sequenced genome from a related species will be an invaluable tool that will facilitate an accurate assessment of the coding potential of a genome (*74*). Studies on the chromosomes of rodent malaria species demonstrated that there was little, if any, difference in gene linkage (*75*), and that large conserved linkage groups could be detected between these rodent malarias and *P. falciparum* (*76*). Limited, but more detailed, analysis revealed that genome organization is highly conserved in the internal non-subtelomeric regions of the chromosome (*77, 78*).

Thus, the expectation is that alignment of orthologous chromosome regions will reveal detailed intron–exon boundaries, and help identify orthologous, but polymorphic, antigen-coding genes, centromeres and species-specific genes. In this last regard, comparison with a more distant apicomplexan genome may also prove valuable. Although gene order may not be well conserved, many salient features of apicomplexans will be revealed through genome comparisons, for example the apicoplast (*79–81*). The ease of transformation of *Toxoplasma gondii* and its suitability for cell biological studies, combined with comparative genomics, provides an excellent opportunity to illuminate malaria biology, develop drug targets (*19, 82*) and possibly vaccines.

Conclusion

The ability to completely sequence genomes has had a huge impact on the way in which scientific research is conducted. As the number of pathogen genomes that have been completed increases, the impetus to develop techniques that utilize this information has increased. In parallel with the *P. falciparum* genome project, techniques such as transfection, microarrays and proteomics, as well as bioinformatic analysis, are

already being developed and applied to fundamental biological questions. The combination of these tools is expected to provide a predictive platform from which to launch hypothesis-driven biological research.

Central to this expectation is continuing financial support for the further development of techniques and resources to exploit the genome sequence. Developments that will provide insight into parasite biology are publicly accessible relational databases, and detailed comparative and survey analyses. It must be emphasized that the overall goal of such costly enterprises is to work towards a cure for malaria, and the contribution of this technological approach could be immense. The current challenge to the research community is to continue the dissemination of this technology and to focus efforts on the development of new drugs and vaccines. ■

Résumé

La recherche sur le paludisme s'engage dans l'ère postgénomique

Le développement des techniques de séquençage du génome a permis de recueillir au cours de ces cinq dernières années une masse d'informations sur ces séquences, qui ont abouti à l'annonce faite récemment qu'on était parvenu à obtenir une première version de travail du génome humain. Les génomes des germes pathogènes, du fait de leur taille réduite, ont été encore plus faciles à analyser avec ces techniques et sont désormais bien représentés dans la génomique. Bien qu'ils ne puissent pas toujours servir de micro-organismes « modèles », l'importance des germes pathogènes en médecine et en agriculture ont fait de l'exploitation des bases de données relatives aux séquences dont ils sont constitués une priorité de tout premier ordre. Mais que signifie réellement la production de ces longues séries d'A-C-G-T sur le plan de la réduction de la charge de morbidité, en particulier dans les pays en développement ? Jusqu'ici, le séquençage du génome a surtout été la spécialité du monde développé, où les ressources et les infrastructures scientifiques en place ont permis la création de centres de séquençage à haut rendement. Toutefois, l'utilisation des résultats de ce séquençage n'a aucune raison d'être ainsi limitée, pour autant qu'on puisse disposer de ressources pour la formation.

L'analyse des données des séquences primaires est probablement l'aspect le plus important de l'ère postgénomique (c'est-à-dire de celle qui va succéder au séquençage proprement dit). C'est ce qu'on appelle la bio-informatique, qui englobe toute une série d'analyses théoriques visant à convertir les séquences d'ADN en informations biologiques. Une part importante de cette discipline a trait à l'identification des gènes par le biais d'un certain nombre de processus, depuis l'analyse des similitudes qu'ils présentent avec des gènes déjà connus d'autres organismes, jusqu'à l'emploi de modèles informatisés basés sur les données existantes. Viennent ensuite les prévisions que l'on peut faire en matière de fonction biologique et de forme moléculaire (génomique structurelle), qui toutes deux ont un potentiel de développement important.

Le fait de connaître l'ensemble des séquences du génome d'un germe pathogène permet également aux chercheurs d'étendre beaucoup plus qu'auparavant leur champ d'investigation du comportement de ces germes. Désormais, au lieu d'étudier l'effet d'un traitement médicamenteux ou la différenciation sur un ou deux gènes, il est possible d'étudier simultanément les variations présentées par l'ensemble des gènes au moyen d'une analyse transcriptionnelle globale. On peut également examiner les protidogrammes, ou modifier génétiquement le micro-organisme (transfection), offrant ainsi un lien direct entre ces effecteurs biologiques et le phénotype macroscopique. La mise au point des techniques nécessaires à ces expériences est une conséquence directe du désir d'exploiter les données du séquençage génomique.

Il n'est pas difficile de prévoir que la possibilité d'identifier et de caractériser le schéma d'organisation génétique des germes pathogènes nous permettra d'identifier les éléments essentiels à leur développement et à leur pathogenèse et de cibler nos efforts de recherche sur la production de nouveaux traitements. En ce qui concerne *Plasmodium falciparum*, on peut identifier les gènes et les protéines qui agissent lors de stades particuliers du cycle évolutif de l'hématozoaire, déterminer leur rôle au moyen de modifications génétiques et utiliser ceux qui sont prometteurs pour la production de vaccins. Les voies métaboliques parasitaires, absentes chez l'hôte, pourraient également servir de cible à des inhibiteurs puissants qui ne soient pas toxiques pour l'homme ; enfin, on pourrait cibler les mécanismes de la pharmacorésistance des plasmodies et allonger ainsi la durée d'efficacité des médicaments existants. Le séquençage du génome de *P. falciparum* offrira de nombreuses possibilités de recherche sur le paludisme, mais il ne s'agit là que d'un début, l'objectif étant de transformer ces possibilités en traitements efficaces sur le terrain.

Resumen

La investigación del paludismo en la era posgenómica

El desarrollo de las tecnologías de secuenciación del genoma a lo largo de los últimos cinco años ha generado una gran cantidad de información sobre secuencias, que ha culminado en el reciente anuncio de un borrador del genoma humano. Debido a su menor tamaño, los genomas de microorganismos patógenos se han prestado aún más a esas metodologías, y actualmente están bien representados en el mundo de la genómica.

Aunque no siempre sirven como microorganismos «modelo», la trascendencia de esos patógenos para la medicina y la agricultura ha hecho que la explotación de las bases de datos de secuencias se convierta en una gran prioridad. ¿Pero cómo puede repercutir realmente la identificación de largas cadenas de A, C, G y T en el objetivo de reducir la carga de morbilidad, en particular en los países en desarrollo? Hasta el momento, la secuenciación de genomas se ha circunscrito en su gran mayoría al mundo desarrollado, donde los recursos y la infraestructura científica han permitido crear centros de secuenciación muy productivos. No obstante, el uso de la información sobre las secuencias no tiene por qué quedar limitado de ese modo, siempre y cuando haya recursos para formación.

El aspecto más importante de la era posgenómica, es decir, la que comienza al completarse la secuenciación, es probablemente el análisis de las secuencias. Este campo, lo que se conoce como bioinformática, comprende una serie de análisis teóricos cuyo fin es convertir la secuencia de ADN en información biológica. Una parte importante de esta disciplina implica la identificación de genes mediante una serie de procedimientos, que van desde la determinación de las similitudes con otros genes descritos anteriormente hasta el uso de modelos informáticos basados en los datos existentes. Posteriormente se hacen predicciones sobre la función biológica y la forma molecular (genómica estructural), cosas ambas que abren amplias perspectivas de desarrollo.

El conocimiento de la secuencia completa del genoma de un patógeno también permite a los científicos investigar el comportamiento de los microorganismos con una base más amplia. Ahora, en lugar de estudiar el efecto de un tratamiento farmacológico o de la diferenciación en uno o dos genes, es posible estudiar la variación simultánea de todos los genes utilizando el análisis transcripcional global. Se puede asimismo examinar la distribución de las proteínas o modificar genéticamente el microorganismo (transfección), y establecer así una relación directa entre esos efectores biológicos y el fenotipo general. La tecnología para llevar a cabo esos experimentos es consecuencia directa del interés por explotar las secuencias genómicas.

Cabe prever que la capacidad para identificar y caracterizar la huella genética de los agentes patógenos ayudará a reconocer elementos clave del desarrollo y los mecanismos patogénicos de los microorganismos causantes de enfermedades, y a centrar nuestras investigaciones en el desarrollo de nuevos tratamientos. En el caso de *Plasmodium falciparum*, es posible identificar los genes y las proteínas que actúan en determinadas etapas del ciclo de vida, verificar su función mediante técnicas de modificación genética, y utilizar candidatos prometedores en la producción de vacunas. Las vías metabólicas del parásito que no están presentes en el huésped son otra posible diana de inhibidores potentes no tóxicos para el ser humano; además, se podría interferir en los mecanismos de farmacorresistencia del parásito, y alargar así la vida eficaz de los medicamentos actuales. La secuencia del genoma de *P. falciparum* brindará muchas oportunidades para investigar el paludismo, pero esto es sólo el principio: el reto es convertir esas oportunidades en tratamientos eficaces sobre el terreno.

References

1. **Triglia T, Kemp DJ.** Large fragments of *Plasmodium falciparum* DNA can be stable when cloned in yeast artificial chromosomes. *Molecular and Biochemical Parasitology*, 1991, **44**: 207–211.
2. **The Wellcome Trust Malaria Genome Mapping Consortium.** The *Plasmodium falciparum* Genome Project: a resource for researchers. *Parasitology Today*, 1995, **11**: 1–4.
3. **Cole ST et al.** Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 1998, **393**: 537–544.
4. **Goffeau A et al.** Life with 6000 genes. *Science*, 1996, **274**: 563–567.
5. **The *C. elegans* Sequencing Consortium.** Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 1998, **282**: 2012–2018.
6. **Hoffman SL et al.** Funding for malaria genome sequencing. *Nature*, 1997, **387**: 647.
7. **Su XZ, Wellems TE.** *Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR. *Experimental Parasitology*, 1999, **91**: 367–369.
8. **Lai Z et al.** A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genetics*, 1999, **23**: 309–313.
9. **Salzberg SL et al.** Interpolated Markov models for eukaryotic gene finding. *Genomics*, 1999, **59**: 24–31.
10. **Pertea M, Salzberg SL, Gardner MJ.** Finding genes in *Plasmodium falciparum. Nature*, 2000, **404**: 34.
11. **Lawson D, Bowman S, Barrell B.** Finding genes in *Plasmodium falciparum. Nature*, 2000, **404**: 34–35.
12. **Gardner MJ et al.** Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum. Science*, 1998, **282**: 1126–1132.
13. **Bowman S et al.** The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum. Nature*, 1999, **400**: 532–538.
14. **Feagin JE.** The extrachromosomal DNAs of apicomplexan parasites. *Annual Review of Microbiology*, 1994, **48**: 81–104.
15. **Wilson RJ, Williamson DH.** Extrachromosomal DNA in the Apicomplexa. *Microbiology and Molecular Biology Reviews*, 1997, **61**: 1–16.
16. **Kohler S et al.** A plastid of probable green algal origin in Apicomplexan parasites. *Science*, 1997, **275**: 1485–1489.
17. **McFadden GI, Roos DS.** Apicomplexan plastids as drug targets. *Trends in Microbiology*, 1999, **7**: 328–333.
18. **Calas M et al.** Antimalarial activity of compounds interfering with *Plasmodium falciparum* phospholipid metabolism: comparison between mono- and bisquaternary ammonium salts. *Journal of Medicinal Chemistry*, 2000, **43**: 505–516.
19. **Jomaa H et al.** Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, 1999, **285**: 1573–1576.
20. **Baruch DI et al.** Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 1995, **82**: 77–87.
21. **Smith JD et al.** Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherence phenotypes of infected erythrocytes. *Cell*, 1995, **82**: 101–110.
22. **Su X-Z et al.** The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum* infected erythrocytes. *Cell*, 1995, **82**: 89–100.

23. **Buffet PA et al.** *Plasmodium falciparum* domain mediating adhesion to chondroitin sulfate A: a receptor for human placental infection. *Proceedings of the National Academy of Sciences of the United States of America*, 1999, **96**: 12743–12748.
24. **Newbold C et al.** Cytoadherence, pathogenesis and the infected red cell surface in *Plasmodium falciparum. International Journal of Parasitology*, 1999, **29**: 927–937.
25. **Carcy B et al.** A large multigene family expressed during the erythrocytic schizogony of *Plasmodium falciparum. Molecular and Biochemical Parasitology*, 1994, **68**: 221–233.
26. **Bonnefoy S et al.** Evidence for distinct prototype sequences within the *Plasmodium falciparum* Pf60 multigene family. *Molecular and Biochemical Parasitology*, 1997, **87**: 1–11.
27. **Bischoff E et al.** A member of the *Plasmodium falciparum* Pf60 multigene family codes for a nuclear protein expressed by readthrough of an internal stop codon. *Molecular Microbiology*, 2000, **35**: 1005–1016.
28. **Weber JL.** Interspersed repetitive DNA from *Plasmodium falciparum. Molecular and Biochemical Parasitology*, 1988, **29**: 117–124.
29. **Cheng Q et al.** *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Molecular and Biochemical Parasitology*, 1998, **97**: 161–176.
30. **Kyes SA et al.** Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America*, 1999, **96**: 9333–9338.
31. **Limpaiboon T et al.** Characterization of a *Plasmodium falciparum* epitope recognized by a monoclonal antibody with broad isolate and species specificity. *Southeast Asian Journal of Tropical Medicine and Public Health*, 1990, **21**: 388–396.
32. **Fernandez V et al.** Small, clonally variant antigens expressed on the surface of the *Plasmodium falciparum*-infected erythrocyte are encoded by the *rif* gene family and are the target of human immune responses. *Journal of Experimental Medicine*, 1999, **190**: 1393–1404.
33. **Kyes S, Pinches R, Newbold C.** A simple RNA analysis method shows *var* and *rif* multigene family expression patterns in *Plasmodium falciparum. Molecular and Biochemical Parasitology*, 2000, **105**: 311–315.
34. **Hayward RE et al.** Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Molecular Microbiology*, 2000, **35**: 6–14.
35. **Eisen MB et al.** Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America,* 1998, **95**: 14863–14868.
36. **Velculescu VE et al.** Serial analysis of gene expression. *Science*, 1995, **270**: 484–487.
37. **Winzeler EA et al.** Direct allelic variation scanning of the yeast genome. *Science*, 1998, **281**: 1194–1197.
38. **Gygi SP et al.** Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 1999, **19**: 1720–1730.
39. **Yates JR, III.** Mass spectrometry. From genomics to proteomics. *Trends in Genetics*, 2000, **16**: 5–8.
40. **Rabilloud T et al.** Analysis of membrane proteins by two-dimensional electrophoresis: comparison of the proteins extracted from normal or *Plasmodium falciparum*-infected erythrocyte ghosts. *Electrophoresis*, 1999, **20**: 3603–3610.
41. **Link AJ et al.** Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 1999, **17**: 676–682.
42. **Uetz P et al.** A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature*, 2000, **403**: 623–627.
43. **Teichmann SA, Chothia C, Gerstein M.** Advances in structural genomics. *Current Opinions in Structural Biology*, 1999, **9**: 390–399.
44. **Tweeddale H, Notley-McRobb L, Ferenci T.** Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool (''pmetabolome'') analysis. *Journal of Bacteriology*, 1998, **180**: 5109–5116.
45. **Kell DB, King RD.** On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends in Biotechnology*, 2000, **18**: 93–98.
46. **Hoffman SL et al.** From genomics to vaccines: malaria as a model system. *Nature Medicine*, 1998, **4**: 1351–1353.
47. **Su X et al.** A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum. Science*, 1999, **286**: 1351–1353.
48. **Goonewardene R et al.** Transfection of the malaria parasite and expression of firefly luciferase. *Proceedings of the National Academy of Sciences of the United States of America*, 1993, **90**: 5234–5236.
49. **Wu Y et al.** Transfection of *Plasmodium falciparum* within human red blood cells. *Proceedings of the National Academy of Sciences of the United States of America*, 1995, **92**: 973–977.
50. **Wu Y, Kirkman LA, Wellems TE.** Transformation of *Plasmodium falciparum* malaria parasites by homologous integration of plasmids that confer resistance to pyrimethamine. *Proceedings of the National Academy of Sciences of the United States of America*, 1996, **93**: 1130–1134.
51. **Crabb BS et al.** Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. *Cell*, 1997, **89**: 287–296.
52. **Crabb BS, Cowman AF.** Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America*, 1996, **93**: 7289–7294.
53. **Horrocks P, Kilbey BJ.** Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. *Molecular and Biochemical Parasitology*, 1996, **82**: 207–215.
54. **Horrocks P, Dechering K, Lanzer M.** Control of gene expression in *Plasmodium falciparum. Molecular & Biochemical Parasitology*, 1998, **95**: 171–181.
55. **Horrocks P, Lanzer M.** Mutational analysis identifies a five base pair cis-acting sequence essential for GBP130 promoter activity in *Plasmodium falciparum. Molecular and Biochemical Parasitology*, 1999, **99**: 77–87.
56. **Dechering KJ et al.** Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum. Molecular and Cellular Biology*, 1999, **19**: 967–978.
57. **Golightly LM et al.** 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum. Molecular and Biochemical Parasitology*, 2000, **105**: 61–70.
58. **Deitsch KW, del Pinal A, Wellems TE.** Intra-cluster recombination and *var* transcription switches in the antigenic variation of *Plasmodium falciparum. Molecular and Biochemical Parasitology*, 1999, **101**: 107–116.
59. **Horrocks P, Lanzer M.** Differences in nucleosomal organization over episomally located plasmids coincides with aberrant promotor activity in *P. falciparum. Parasitology International*, 1999, **48**: 55–61.
60. **Lobo CA et al.** Disruption of the Pfg27 locus by homologous recombination leads to loss of the sexual phenotype in *P. falciparum. Molecular Cell*, 1999, **3**: 793–798.
61. **Triglia T et al.** Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *EMBO Journal*, 1998, **17**: 3807–3815.
62. **Reed MB et al.** Pgh1 modulates sensitivity and resistance to multiple antimalarials in *Plasmodium falciparum. Nature*, 2000, **403**: 906–909.

63. **Craig AG, Waters AP, Ridley RG.** Malaria Genome Project Task Force – A post-genomic agenda for functional analysis. *Parasitology Today*, 1999, **15**: 211–214.
64. **Waters AP et al.** Transfection of malaria parasites. *Methods*, 1997, **13**: 134–147.
65. **van Dijk MR, Waters AP, Janse CJ.** Stable transfection of malaria parasite blood stages. *Science*, 1995, **268**: 1358–1362.
66. **Menard R et al.** Circumsporozoite protein is required for development of malaria sporozoites in mosquitoes. *Nature*, 1997, **385**: 336–340.
67. **Sultan AA et al.** TRAP is necessary for gliding motility and infectivity of *Plasmodium* sporozoites. *Cell*, 1997, **90**: 511–522.
68. **van der Wel AM et al.** Transfection of the primate malaria parasite *Plasmodium knowlesi* using entirely heterologous constructs. *Journal of Experimental Medicine*, 1997, **185**: 1499–1503.
69. **Kochen C, van der Wel A, Thomas AW.** *Plasmodium cynomolgi*: Transfection of blood stage parasites using heterologous DNA constructs. *Experimental Parasitology*, 1999, **93**: 58–60.
70. **Roos DS et al.** Transport and trafficking: *Toxoplasma* as a model for *Plasmodium. Novartis Foundation Symposium*, 1999, **226**: 176–195.
71. **Soete M, Hettman C, Soldati D.** The importance of reverse genetics in determining gene function in apicomplexan parasites. *Parasitology*, 1999, **118**: S53–61.
72. **Roos DS et al.** Origin, targeting, and function of the apicomplexan plastid. *Current Opinions in Microbiology*, 1999, **2**: 426–432.
73. **Beverley SM, Turco SJ.** Lipophosphoglycan (LPG) and the identification of virulence genes in the protozoan parasite Leishmania. *Trends in Microbiology*, 1998, **6**: 35–40.
74. **Rubin GM et al.** Comparative genomics of the eukaryotes. *Science*, 2000, **287**: 2204–2215.
75. **Janse CJ et al.** Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. *Molecular and Biochemical Parasitology*, 1994, **68**: 285–296.
76. **Carlton JM et al.** Gene synteny in species of *Plasmodium. Molecular and Biochemical Parasitology*, 1998, **93**: 285–294.
77. **Vinkenoog R et al.** Malaria parasites contain two identical copies of an elongation factor 1 alpha gene. *Molecular and Biochemical Parasitology*, 1998, **94**: 1–12.
78. **van Lin LHM, Janse CJ, Waters AP.** The conserved genome organization of non-falciparum malaria species: the need to know more. *International Journal of Parasitology*, 2000, **30**: 357–370.
79. **Waller RF et al.** Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America*, 1998, **95**: 12352–12357.
80. **Waller RF et al.** Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *EMBO Journal*, 2000, **19**: 1794–1802.
81. **Striepen B et al.** Expression, selection, and organellar targeting of the green fluorescent protein in *Toxoplasma gondii. Molecular and Biochemical Parasitology*, 1998, **92**: 325–338.
82. **Roos DS.** The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites: some additional thoughts. *Parasitology Today*, 1999, **15**: 41.
83. **Dessens JT et al.** CTRP is essential for mosquito infection by malaria ookinetes. *EMBO Journal*, 1999, **18**: 6221–6227.
84. **Yuda M, Sakaida H, Chinzei Y.** Targeted disruption of the *Plasmodium berghei* CTRP gene reveals its essential role in malaria infection of the vector mosquito. *Journal of Experimental Medicine*, 1999, **190**: 1711–1716.
85. **Templeton TJ, Kaslow DC, Fidock DA.** Developmental arrest of the human malaria parasite *Plasmodium falciparum* within the mosquito midgut via CTRP gene disruption. *Molecular Microbiology*, 2000, **36**: 1–9.
86. **Wengelnik K et al.** The A-domain and the thrombospondin-related motif of *Plasmodium falciparum* TRAP are implicated in the invasion process of mosquito salivary glands. *EMBO Journal*, 1999, **18**: 5195–5204.
87. **Kappe S et al.** Conservation of a gliding motility and cell invasion machinery in Apicomplexan parasites. *Journal of Cell Biology*, 1999, **147**: 937–944.
88. **O'Donnell RA et al.** Functional conservation of the malaria vaccine antigen MSP-119 across distantly related *Plasmodium* species. *Nature Medicine*, 2000, **6**: 91–95.