

Showcase on Research

Bioinformatics and the Malaria Genome

Ross L. Coppel and Casilda G. Black

Department of Microbiology and the Victorian Bioinformatics Consortium, Monash University, Clayton, VIC 3800

Introduction

Malaria infection by protozoan parasites of the genus *Plasmodium* is one of the most important human infections. Such infections account for more than 300 million clinical cases, resulting in more than two million deaths every year. This is about 4% of yearly human mortality.

The majority of the toll of malaria is due to infection by the species *Plasmodium falciparum*. However, relatively little funding is devoted to the study of this parasite and malariology has languished as a cottage industry for many years. Nevertheless, the face of biology is changing and the driving force is genomics.

Even in the sleepy backwaters of tropical medicine, the vivifying force of a complete gene sequence is changing what scientists do. They are having to come to grips with the enormous amount of data generated by the genome and various functional genomics projects such as microarrays and proteomics.

Various bioinformatics strategies are being developed to provide new ways of storing and analysing all this data, in order to facilitate the discovery of new knowledge and the development of new treatments and diagnostic tests.

History of the genome project

The genome of *P. falciparum* is between 25 and 30 Mb in length and contains 14 chromosomes ranging in size from 0.6 to 3.5 Mb (1). The DNA is remarkably AT-rich with an average GC content of 18% (2), a property believed to account for the instability of large fragments of *P. falciparum* DNA in *Escherichia coli* (3). The *P. falciparum* genome project was initiated subsequent to the completion of a successful chromosome mapping project funded by the Wellcome Trust. By 1996 scientists were actively debating the feasibility of sequencing the entire genome of *P. falciparum*. An international consortium was established comprising three sequencing centres, The Institute for Genomic Research (TIGR, Rockville, MD, USA), the Sanger Centre (Hinxton, UK) and Stanford University (Stanford, CA,

USA). Funding was supplied by the Burroughs Wellcome Fund, the National Institutes of Health and the U.S. Department of Defense. The purpose of the consortium was to completely sequence the genome and to annotate it, with a target completion date of 2002–2003. The 14 chromosomes were divided between the three centres and sequencing began (4,5).

The approach to the genome was dictated in part by the split nature of the project. Chromosomes of *P. falciparum* are not rendered visible by conventional karyotypic procedures, so the chromosomes were separated by pulsed field gel electrophoresis and purified from the gel. These individual chromosomes were sheared and used to construct chromosome-specific libraries.

Shotgun sequencing of these libraries was performed and gap closure undertaken by the usual mix of methods including PCR and pinning to the YAC contigs (i.e. overlapping plasmodium DNA segments cloned in yeast artificial chromosomes) determined during the mapping project. Not all chromosomes could be separated and at least 3 chromosomes co-migrate in a structure euphonyously named the BLOB. This is being sequenced as a unit and, of course, the sequence will be a more complex project to assemble.

At the time of its inception in the mid-1990s, it was not clear whether the sequencing of an organism with such an AT-rich genome was technically feasible, given the numerous reports of difficulties that had bedevilled the sequencing of individual genes (3). The success in finishing chromosomes 2 and 3 laid these fears partly to rest, however some gaps in the sequence are proving quite difficult to close, and some chromosomes started at about the same time as 2 and 3, such as 1 and 4, are still not completed. Accordingly, the sequencing consortium is adopting a similar process to the human genome project and producing a "first draft" of the genome that will be largely completed by the end of 2001. It is thought that essentially all genes will be present, at least in part, although whether they are recog-

nised as such depends on the success of the annotation process (see article by Robert Huestis on page 17 of this issue).

Current resources

Chromosomes 2 and 3 have been completed and reported. Chromosome 2 is 947 kb in length and contains 210 predicted genes, whereas chromosome 3 is 1,060 kb and contains 215 predicted genes. The sub-telomeric regions of both chromosomes revealed a conserved order of features, including repetitive DNA sequences, members of multigene families involved in virulence and antigenic variation, a number of conserved pseudogenes, and several genes of unknown function.

Genes in this area included the variant antigen genes (*var*) that encode the PfEMP1 proteins (6), and members of the *rif* and *stevor* gene families (7). The *rif*ins, which encode putative surface proteins, appear to comprise up to 7% of the protein-encoding gene in the genome. Many genes on these chromosomes showed signs of a plastid or mitochondrial origin, including several genes involved in fatty acid biosynthesis.

Collectively these 2 chromosomes revealed that gene density is of the order of one gene about every 5 kb and this would give an estimate of about 6000 to 7000 genes in the genome. This is a lower gene density than, say, *Saccharomyces cerevisiae* but higher than human. In contrast to yeast, there are a greater proportion of genes containing introns and nearly twice as many proteins containing predicted non-globular domains. Almost half of the genes are predicted to contain at least one intron; however, some introns will have been missed by the gene-finding programs.

Already the first studies that build on the genome sequence to augment our arsenal of anti-malarial strategies are appearing (8). Recognition of enzymes involved in the mevalonate-independent pathway of isoprenoid synthesis within the genome data enabled scientists to focus in on a potentially vulnerable pathway. Indeed, inhibitors of this pathway have already

Showcase on Research

Bioinformatics and the Malaria Genome (contin.)

been found and two compounds have been identified that block the pathway and cure mice of experimental malaria infection. Scanning of the genome data has also revealed potential candidate vaccine molecules, including another merozoite surface protein that contains two epidermal growth factor-like domains, a structure similar to that found in MSP1 (9).

There are several repositories of genomic sequence information and a number of these are listed in **Table I**. The primary sites are, of course, the sequencing centres at TIGR, the Sanger Centre and Stanford University. These sites concentrate on the data generated at each respective centre and are supplemented by varying degrees of analysis and annotation, both manual and automated.

Chromosomes 2 and 3 have been completed (10,11), while sequencing and annotation of the remaining chromosomes is currently at various stages of completion. An obstacle encountered by the sequencing centres has been that the gene-finding algorithms that are utilised for other genomes such as fruit fly or human are less successful when confronted with the high AT-rich sequence of *P. falciparum* in particular. Retraining of the gene finders on malaria sequence has yielded improvements, such as the improved performance of GlimmerM over Glimmer (12).

Although the sequencing centres view these models as provisional, there is a tendency among scientists to accept them as correct, and base their experiments on this information. Thus, a search for genes encoding exported proteins in asexual stages, looking for genes that have a 2-exon structure, would have examined a fairly large number of false positives based on the initial chromosome annotations, as well as missing some 2-exon genes not properly identified. It must be remembered that the current annotations are predictions and must be verified by experiment. Some *P. falciparum* genes can be quite unusual in structure, being composed of multiple small exons; current prediction software do not easily predict these genes.

How do we make genome data more accessible?

Effective utilisation of the genome sequence will be the ultimate criterion of

success for the project and so it becomes important to ensure that useability and accessibility of the sequence data are enhanced. What factors militate against full exploitation of the genome?

There are a number of factors specific to the nature of the malaria field itself, but a problem common to many groups is the relative newness of information of this sort and the relative paucity of bioinformatics expertise in the scientific community. While just about everyone knows how to do a BLAST search on some web site, most lack the skills to find the rest of the treasure buried in this information.

Bioinformatics analysis requires a particular set of skills and a knowledge base that has not traditionally been part of standard biological teaching and practice. Although new courses in Bioinformatics are appearing, few practicing scientists have other than a passing familiarity with this area. It may be difficult for scientists unskilled in this area to fully appreciate how the genome may be valuable in advancing their own field of interest.

Beyond the BLAST search are a number of sophisticated analyses that can give insights into parasite metabolism or phylogeny that are of great value. This problem can be ameliorated to some extent by collaborations with specifically trained bioinformaticians who can contribute their skills to a shared project.

Several factors mandate the need for improved databases. The fragmented nature of the genome project itself and the existence of individual sequences at the three sequencing centres absolutely required the design of a single repository for those sequences in the longer term.

Once the centres complete their sequencing tasks, it is unlikely that significant resources will be allocated to maintenance and ongoing curation and annotation of this particular genome. However, it is essential that the genome continue to be annotated over time, as experimental verification will be required for many gene models. The results will need to be incorporated into the databases in a timely manner, as will the results from many other forms of experimentation. This responsibility will fall to the malaria community, which must support ongoing development and curation of the genome database. The study of malaria did not commence with the genome sequence.

There is a considerable amount of important information already available on various malaria proteins. It is important that this information be integrated with the genome sequence, and functional genomics results. Similarly, many reagents of various sorts have been generated as a result of these experiments, and appropriate pointers to such reagents should be built into the database.

Finally, *P. falciparum* is only the first malaria species to be sequenced. There are projects that are in various stages of planning and execution to sequence at least an additional 6 malaria species and strains. In the near future, we will see attempts at improving these databases, so that information is more readily obtained and the types of queries that can be readily answered made more powerful. This is not a problem for malariologists alone, and is being faced by all scientists coming to grips with this next phase of biology.

Thus it is likely that there will be new developments in databases in other systems that can be usefully applied to malaria. In malaria, the first attempts at such an integrated database is occurring under the auspices of the PlasmoDB project. This is project funded by the Burroughs Wellcome Fund involving scientists at Monash University and the University of Pennsylvania and is a project to develop an organism-specific genome database for *Plasmodium* spp., that integrates genomic and functional genomics data with all the rest of malariology. Interested readers can follow its development at the PlasmoDB web site listed in **Table I**.

Finally, it is worth noting that for a country in which malaria is not endemic, Australia has a surprisingly strong research presence in malaria. Internationally recognised groups are located throughout Australia, in Melbourne at the Walter and Eliza Hall Institute, Monash University, the University of Melbourne and La Trobe University, in Queensland at the Queensland Institute of Medical Research, in Sydney at the University of New South Wales and elsewhere, including the Australian National University. This is not an exhaustive list by any means. With such a strong background, it is not surprising that Australia has contributed to aspects of the genome project.

We have mentioned the malaria genome mapping project, which was the

Showcase on Research

Bioinformatics and the Malaria Genome (contin.)

prelude to the full scale sequencing project and provided YAC contig maps which aided sequence assembly. This project was initiated by David Kemp while he was at the Walter and Eliza Hall Institute, and operated as an international collaborative network, a model later used for the genome project proper. There have been important contributions in the area of bioinformatics and sequence databases with the WHO TDR malaria sequence database, curated by one of the authors, and the PlasmoDB project.

Acknowledgements

RLC is supported by the Victorian Bioinformatics Consortium, the Australian National Health and Medical Research Council, the UNDP/World Bank/WHO

Special Programme for Research and Training in Tropical Diseases, the Howard Hughes Medical Institute International Scholars in Infectious Diseases and Parasitology Program and the Burroughs Wellcome Fund.

References

1. Triglia, T., Wellem, T. E., and Kemp, D. J. (1992) *Parasitol. Today* **8**, 225-229
2. Pollack, Y., Katzen, A. L., Spira, D. T., and Golenser, J. (1982) *Nucleic Acids Res.* **10**, 539-546
3. Coppel, R. L., and Black, C. G. (1998) in *Malaria: Parasite biology, pathogenesis and protection* (Sherman, I. W., ed), pp. 185-202, ASM Press, New York.
4. Carucci, D. J., Gardner, M. J., Tettelin, H., Cummings, L. M., Smith, H. O., Adams, M. D., Hoffman, S. L., and Venter, J. C. (1998) *Exp. Rev. Mol. Med.* **5 May**, <http://www-ermm.cbccu.cam.ac.uk/dcn/txt001dcn.htm>
5. Gardner, M. J. (1999) *Curr. Opin. Genet. Dev.* **9**, 704-708
6. Su, X. Z., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., Ravetch, J. A., and

7. Cheng, Q., Cloonan, N., Fischer, K., Thompson, J., Waite, G., Lanzer, M., and Saul, A. (1998) *Mol. Biochem. Parasitol.* **97**, 161-176
8. Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H. K., Soldati, D., and Beck, E. (1999) *Science* **285**, 1573-1576
9. Black, C. G., Wu, T., Wang, L., Hibbs, A. R., and Coppel, R. L. (2001) *Mol. Biochem. Parasitol.* **114**, 217-226
10. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Smith, H. O., Fraser, C. M., Venter, J. C., and Hoffman, S. L. (1999) *Parassitologia* **41**, 69-75
11. Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T., Gentles, S., Gwilliam, R., Hamlin, N., Harris, D., Holroyd, S., Hornsby, T., Horrocks, P., Jagels, K., Jassal, B., Kyes, S., McLean, J., Moule, S., Mungall, K., Murphy, L., Barrell, B. G., and et al. (1999) *Nature* **400**, 532-538
12. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shalloom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J. P., Aston, C., Lai, Z. W., Schwartz, D. C., Perlea, M., Salzberg, S., Zhou, L. X., Sutton, G. G., Clayton, R., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Hoffman, S. L., et al. (1998) *Science* **282**, 1126-1132

Table 1. Malaria genome resources available on the web

Web Resource	URL	Features	Comments
Sequencing Centres			
Sanger Centre	http://www.sanger.ac.uk/Projects/P_falciparum/	BLAST, Search of annotated genes	Focus is portion of the genome sequenced at Sanger
TIGR	http://www.tigr.org/tdb/edb2/pfal/htmls/	BLAST, Search of annotated genes	Focus is portion of the genome sequenced at TIGR
Stanford University	http://sequence-www.stanford.edu/group/malaria/index.html	BLAST, Sequence retrieval and search of annotated genes	Chromosome 12 only. Community annotation of this chromosome may be done here.
Gene Sequence Tag Project	http://parasite.vetmed.ufl.edu/	<i>P. berghei</i> and <i>P. vivax</i> gene sequencing projects	Useful early information on these genomes. Forthcoming genome projects will provide more complete data elsewhere
Malaria Full Length cDNA project	http://fullmal.ims.u-tokyo.ac.jp/	Sequencing runs taken from a library of putative full length cDNA clones	Needs further data before it becomes truly useful
Data Analysis Sites			
NCBI Malaria Genetics and Genomics	http://www.ncbi.nlm.nih.gov/Malaria	BLAST, microsatellites, genome data collection, optical mapping, ESTs	Tool to map candidate genes using genetic cross
WHO TDR Malaria Database	http://www.who.edu.au/MalDB-www/who.html	Genome data collection, bibliographies, sequence alignments, SRS query engine, malaria discussion group	Annotated sequences may be searched in a nomenclature independent manner. Best site to collect multiple alleles of a single gene
PlasmoDB	http://www.plasmodb.org	BLAST, chromosome maps	Under continuous development to meet community needs. Should become the most useful genome site
MR4	http://www.malaria.mr4.org/mr4pages/index.html	Access to research reagents, BLAST and sequence retrieval facility planned	Malaria Research Protocols available
Malaria Parasite Metabolic Pathways	http://sites.huji.ac.il/malaria/	KEGG style pathway maps and useful bibliography, links to Expasy	Not linked to sequence data at present. Some attempts to look at more complex phenomena such as transport, cytoadherence and rosetting
SANBI Plasmodium	http://www.sanbi.ac.za/malaria-genesearch/	Assemblies and BLAST comparisons of sequences from <i>P. falciparum</i> , <i>P. vivax</i> and <i>P. berghei</i>	In its present incarnation, it is difficult to navigate and hard to extract information. Likely to become more useful when genome data is completely available
Parasite Proteome Server	http://www.ebi.ac.uk/parasites/proteomes.html	Classification of proteins of <i>Plasmodium falciparum</i> into functional categories	Attempt at placing genes in functional categories. Minimally curated. Will be more useful as genome project is completed
Pedant Analysis of Chromosome 2	http://pedant.mips.biochem.mpg.de/cgi-bin/wwwfly.pl?Set=PfalciparumI&Page=index	Complete automated annotation of 205 proteins on chromosome 2	Riddled with errors due to the automated nature of the process, including mistakes in localisation and similarity.

This is not an exhaustive listing but describes some features of the more important sequence and genome resources available. Many are being continuously updated and features will change with time.