

The grand assault

Russell F. Doolittle

The complete genome sequence of the parasite responsible for most of the world's human malaria has been determined. The nature of the genome meant that this was a difficult project, requiring considerable ingenuity.

The parasite *Plasmodium falciparum*, responsible for most human malaria, is among the most studied pathogens of all time, probably surpassed only by the human immunodeficiency virus and the tuberculosis bacterium *Mycobacterium tuberculosis*. The extent of human suffering caused by malaria and its devastating costs have long been recognized by international bodies, and many initiatives have been taken over the years to try to defeat this insidious microbe¹. In 1996, an international consortium of scientists from more than a dozen institutions set out to determine the 23 million base pairs of DNA that make up the organism's genome sequence. Their massive effort — which ended up going well beyond simple sequencing — is reported on pages 498–542 of this issue^{2–8}. The avowed goal of the project was to search for chinks in the parasite's armour, so that new and effective drugs and vaccines might be developed.

Sequencing strategy

The strategy for determining the *P. falciparum* genome sequence depended on first physically separating its 14 chromosomes by the technique of pulsed gel electrophoresis. In fact, three of the chromosomes (numbers 6, 7 and 8) could not be separated from each other and were simply taken as a combined unit. Three different teams then attacked different chromosomes: a team led by the Sanger Centre, Cambridge, UK, sequenced nine³; The Institute for Genomic Research (TIGR), Maryland, and others took on four⁴; and a group centred on Stanford University, California, did the other⁵.

In broadest outline, the DNA was mechanically sheared into random fragments, the fragments were inserted into bacteria (where they were copied every time the bacteria multiplied), and individual bacterial colonies were collected. The DNA inserts from these clones were sequenced automatically, and their order determined by assembling overlapping sequences together using a computer. Around half a million individual fragments were sequenced. As in most genome projects on this scale, the sequence determination is not really complete, and several gaps and ambiguities remain. Nonetheless, even at 95% 'finished', the published sequence must be regarded as a milestone that will be a major asset to biomedical researchers.

But why did this project take so long? After all, 23 million base pairs is not a particularly large genome by current standards (Fig. 1, overleaf), and the much larger genome of the fruitfly *Drosophila melanogaster* was apparently completed in less than a year⁹. The single biggest hurdle was the extremely biased base composition of the *P. falciparum* genome. More than 80% of the bases are either As or Ts (as opposed to Cs or Gs). In fact, regions of the genome that do not code for genes average more than 85% As or Ts, and runs of 50 As or Ts are common. Most of the genomes

already sequenced have been much less skewed in their base composition.

The extreme bias made the assembly process — by which individual clones are put in their correct order by an iterative overlap process — particularly challenging. Usually, if one clone has a distinctive sequence at one end (say, its 38 end) and another the exact same sequence at the other (58) end, it is assumed that these sequences overlap in the genome. But for *P. falciparum*, so many clone ends were AT-rich that it was difficult to assign overlaps. As a result, new stratagems had to be devised for ordering many of the chromosomal pieces, including a heavy reliance on genetic and physical 'maps' of genomic landmarks.

For example, one type of physical map used was an 'optical' map. Here, a purified chromosome is cut into segments with an enzyme known to cut DNA at particular sequences, and the segments are separated according to size by gel electrophoresis, producing the optical map. Meanwhile, the postulated sequence is 'virtually' fragmented in a computer by breaking it at the theoretical sites at which the chosen enzyme cuts. The hypothetical fragments are then sorted by length, generating a virtual map. The agreement between the optical and virtual maps for most chromosomes was reassuringly good³.

Identifying the genes

Extreme AT-richness aside, finding the genes in any eukaryotic genome can be problematic, because the protein-coding parts of genes (exons) are interrupted by non-coding regions (introns). (Eukaryotes can be loosely defined as organisms whose cells have nuclei and cytoskeletons, distinguishing them from the Bacteria and the Archaea — neither of which has introns in their coding sequences.) Although computer programs can identify the ends of exons that need to be joined together to form mature gene products, these are seldom 100% accurate. So there is often a need for validation that is not required in bacterial gene analysis. Such confirmation can be provided by studies of complementary DNA sequences, which directly reflect the mature gene products, or by identification of the encoded proteins themselves^{6,7}. To achieve the latter, the consortium used ultra-sensitive mass spectrometry — the application of which is almost sure to become a standard component of future genome projects of this sort.

Frustratingly, possible functions for fully 60% of the postulated 5,279 genes remain unknown, because these

This week's News Features section has two further articles describing reaction to publication of the *Plasmodium* genomes, and discussion of the prospects for malaria control. See pages 426 and 429.

news and views

genes match no other sequences in existing data banks. Another 5% of the genes are also classified as ‘hypothetical’ in this sense, although they do have counterparts — themselves with unknown functions — in other organisms. This is both surprising and disappointing. But we can be sure that many of these genes really do exist. For instance, mass spectrometry identified authentic peptides corresponding to proteins encoded by 2,391 of the genes, including many of those for which functions have not yet been found^{6,7}.

A second *Plasmodium* sequence

Remarkably, the consortium also sequenced a second plasmodial genome — that of *P. yoelii yoelii*⁸, the cause of a malaria-like disease in wild African rats. More than 60% of the *P. falciparum* genes, including most general housekeeping genes, had close relatives in the *P. yoelii yoelii* genome. But the comparison also revealed a treasure-trove of differences and rearrangements. Many of these are near the ends of chromosomes, in regions that somehow control the impressive ability of plasmodial parasites to change and thereby evade recognition by the host immune system. There is evidence in both species for the kinds of genetic rearrangements and chromosomal exchanges that might be allied to this ability. But none of the genes known to be involved in immune evasion by *P. falciparum* can be recognized in *P. yoelii yoelii*. The hope had been that comparison would reveal host-specific adaptations that could be exploited in some way, but the extreme differences have confounded that strategy.

The hunt for vulnerability

Having the entire inventory of genes for *P. falciparum* provides a complete map of its metabolic pathways, the genes encoding metabolic enzymes all being recognized by comparison with other organisms. Not only can the pathways active at different stages of the parasite life cycle be delineated, but key points at which the organism’s metabolism is known to be vulnerable to attack can be seen in an overall context. For example, quinine — the first and most successful antimalarial drug — acts within a subcellular compartment of the parasite, the food vacuole, in which host haemoglobin is degraded as a foodstuff. Sulphonamides are effective because *P. falciparum* makes its own folic acid vitamins by a scheme involving *p*-aminobenzoic acid — a structure mimicked by sulphanilamide. At least four other known targets of antimalarial drugs reside in the apicoplast, an exotic quadruple-membraned compartment.

Most of these sites of drug action were known well before the genome-inspired metabolic map, although it is reassuring to see the whole landscape. It is the potential for

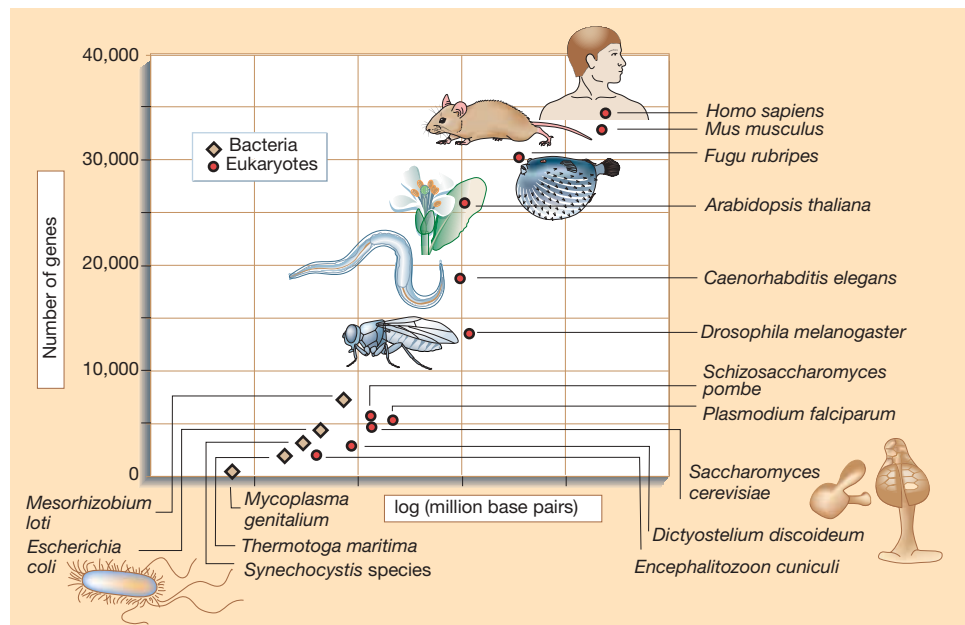


Figure 1 Some of the genomes sequenced so far. The figure shows the number of genes plotted against genome size for the 12 fully sequenced genomes of eukaryotes and a representative set of bacteria. Note the log scale for genome size, expressed as millions of base pairs.

choosing new drug targets that is exciting, however, and the consortium has now pinpointed five. For example, within the food vacuole there are several protein-degrading enzymes that might conceivably be blocked by specific inhibitors. How long it will be before these hopes are realized is unknown, but there are almost too many options to pursue. The question also arises of whether drugs that might be forthcoming would be affordable by those most in need.

Weighing the costs

During the course of this project there has been a spirited debate as to whether malaria is better attacked by large-scale genome projects¹⁰ or by more traditional public-health measures¹¹. The politically correct chorus of response has been that of course both approaches must be undertaken, especially as the true benefits of the genome studies are “down the line”^{12,13}.

Initially, the cost of the current project was put in the neighbourhood of US\$15 million¹⁴. These are not easy estimates to make because funding usually comes from multiple sources and sometimes involves third parties. To the initial figure should now be added the cost of many parallel projects, including the now published sequence¹⁵ of *Anopheles gambiae*, the mosquito that transmits malaria. The Sanger Centre and TIGR websites also list half a dozen other protozoan parasite genomes under study, including two more plasmodial strains.

So is it worth it from a medical point of view? That really remains to be seen. If one assesses what has been learned so far from the whole-genome projects (see Fig. 1) of the past six or seven years, it is clear that the

‘bio’ part of the biomedical enterprise has been the clear winner. Whether one studies molecular evolution or gene transcription, population genetics or developmental biology, cellular mechanics or signal transduction, whole-genome information is what defines the playing field. But for the most part, the promised medical benefits have been slow to materialize. Translating all of this information into new treatments and cures is not a trivial process.

That malaria was near eradication decades ago in some areas as a result of DDT spraying is more than a cruel irony^{11,16}. Indeed, since the use of that insecticide was sharply curtailed in the 1970s, 30 million people may have died from *Plasmodium*-infected mosquito bites, and ten times that number have suffered this debilitating disease. As the participants in the current study acknowledge², “genome sequences alone provide little relief to those suffering from malaria”.

Russell F. Doolittle is at the Center for Molecular Genetics, University of California, San Diego, La Jolla, California 92093-0634, USA.
e-mail: rdoolittle@ucsd.edu

1. Malaria Insight *Nature* **415**, 669–715 (2002).
2. Gardner, M. J. *et al.* *Nature* **419**, 498–511 (2002).
3. Hall, N. *et al.* *Nature* **419**, 527–531 (2002).
4. Gardner, M. J. *et al.* *Nature* **419**, 531–534 (2002).
5. Hyman, R. W. *et al.* *Nature* **419**, 534–537 (2002).
6. Florens, L. *et al.* *Nature* **419**, 520–526 (2002).
7. Lasonder, E. *et al.* *Nature* **419**, 537–542 (2002).
8. Carlton, J. M. *et al.* *Nature* **419**, 512–519 (2002).
9. Adams, M. D. *et al.* *Science* **287**, 2185–2195 (2000).
10. Hoffmann, S. L. *Science* **290**, 1509 (2000).
11. Curtis, C. F. *Science* **290**, 1508 (2000).
12. James, A. A. *Science* **291**, 435–436 (2001).
13. Morel, C. M. *Science* **291**, 435–436 (2001).
14. Hoffmann, S. L. *et al.* *Nature* **387**, 647 (1997).
15. Holt, R. A. *et al.* *Science* **298**, 129–149 (2002).
16. Jukes, T. H. *Am. Sci.* **51**, 355–361 (1965).