From: Malaria Parasites: Genomes and Molecular Biology
Edited by: A.P. Waters and C.J. Janse

Chapter 3

Getting the Most Out of Bioinformatics Resources

Jessica C. Kissinger
and David S. Roos

Abstract

The recent publication of a complete reference sequence for the *Plasmodium falciparum* genome is a momentous event for malaria researchers. In addition, genomic and functional genomics data is now available for six further *Plasmodium* species and eight non-*Plasmodium* species of apicomplexan parasites. These datasets can greatly expedite the identification of candidate targets for drug, vaccine and diagnostic development, in addition to enhancing our basic understanding of malaria parasites. But how can researchers most effectively access and exploit genomic-scale data, integrating this information with the results from other experiments? Bioinformatics research is fundamentally no different from 'wet lab' experiments conducted at the bench, requiring an understanding of the starting reagents (databases), the strengths and weaknesses of experimental (computational) methods, and a critical analysis of the results obtained. This chapter discusses the nature and organization of data resources, strategies for data mining, and the interpretation of computational results.

1. Introduction

Now is an exciting time to be engaged in malaria research. Significant technological advances, research effort, and financial investment have produced a complete reference genome for *Plasmodium falciparum* (Gardner *et al.*, 2002b), effectively complete genome sequences for *P. yoelii, knowlesi, vivax,* and several other apicomplexan parasite species (see below), and complete genome sequences for both the human (and mouse) host and the *Anopheles gambiae* vector (Holt *et al.*, 2002). Advances in genomics and bioinformatics affect all malaria researchers, from the molecular biologist interested in rifin gene organization, to the developmental biologist studying stage-specific gene expression, to the cell biologist investigating protein trafficking to Maurer's clefts, to the evolutionary biologist exploring the origins of cytoadherence ligands, to the immunologist seeking a target for vaccine development, to the population geneticist studying allelic variation for evidence of positive or negative selection. How can we as parasitologists access, manage, and utilize the surfeit of emerging data, integrating 'dry-lab' computational research with 'wet-lab' studies at the laboratory bench?

The search for – and functional analysis of – genes is increasingly moving towards high-throughput studies of the whole genome in parallel, generating huge data sets related to transcript expression and transcriptional regulation, protein translation and steady-state levels, protein-protein interactions, polymorphic diversity, etc. All of these data need to be stored, analyzed, and made widely accessible. Many useful datasets and analysis tools are now accessible on-line, offering great potential for expediting malaria research and stimulating new lines of investigation, but these resources are scattered throughout the internet. One purpose of this chapter is to provide a compendium of on-line resources for the malaria researcher.

Of course, the development of new kinds of data also brings the need for new tools and skills to navigate the continually changing information landscape, whether to find the 5′ end of your favorite gene, determine the hypothetical function for cDNAs on a microarray, or identify a potential target for drug, vaccine, or diagnostic development. Indeed, it is increasingly possible – and at times even essential – to conduct malaria research *in silico*. While this does not obviate the need for wet-lab experimentation, computational approaches often provide a useful complement, and can be much faster than conventional benchwork; the effective integration of wet-lab and computational approaches can produce dramatic research advances. A second purpose of this chapter is to provide a bioinformatics primer for malaria researchers.

Bioinformatics for the average bench scientist and expert alike has evolved rapidly over the last decade, and need not be a "black box". The inner workings of the many common tools are well described in various

textbooks (Baxevanis and Ouellette, 2001; Gibson and Muse, 2001; Mount, 2001; Baxevanis *et al.*, 2003) and will not be discussed here in any detail. The field has also accumulated experience with the efficient application of bioinformatics tools and approaches. Bioinformatics techniques, like laboratory techniques, can generate misleading results, and it is therefore critical to understand the strengths and weaknesses of the methods employed. A properly designed bioinformatics experiment includes controls and safeguards designed to detect potential artifacts.

Other contributions to this book describe laboratory approaches and techniques now being applied to *Plasmodium*, many of which employ and/or generate genomic-scale data. In this chapter, we take a 'behind the scenes' look at how genomic data are generated, stored and analyzed, in hopes that this will enable researchers to design experiments that use these data effectively. We begin with a brief introduction to the types of data that one can expect to find, including some relevant information on how these data are processed. The following section describes currently available *Plasmodium* resources, with notes on data organization. We then discuss the power of database queries, illustrated by a series of queries intended to elucidate candidate vaccine antigens (with suggestions for effective data-mining and notes on common pitfalls). Finally, we provide a few notes on newly emerging data types, and opportunities for the future.

2. The Datasets

Genomic-scale datasets are currently available for several *Plasmodium* species, including *P. berghei, chabaudi, falciparum, knowlesi, reichenowi, vivax, yoelii*. Table 1 indicates which data types are publicly available for each species as of October 2003. A large amount of data is also available for other related apicomplexan species, including *Babesia bovis, Theileria annulata, Theileria parva, Toxoplasma gondii, Eimeria tenella, Neospora caninum, Sarcocystis neurona* and *Cryptosporidium parvum* (Table 2), and for the human and mouse hosts and the insect vector *Anopheles gambiae* (Table 3).

Data types available for *Plasmodium falciparum* include information on:

Nucleotide Sequences

- Genomic sequence (del Portillo *et al.*, 2001; Tchavtchitch *et al.*, 2001; Carlton *et al.*, 2002; Gardner *et al.*, 2002a) and Genome Survey Sequences (GSS) (Carlton and Dame, 2000; Janssen *et al.*, 2001).
- Expressed Sequence Tags (EST) (Watanabe *et al.*, 2001; Li *et al.*, 2003).

RNA Expression

- ESTs and Serial Analysis of Gene Expression (SAGE) tags (Munasinghe *et al.*, 2001)

Table 1. Data sources for *Plasmodium* and other Apicomplexan parasite species

Species	Data type(s)	URL
All *Plasmodium* species	multiple	http: //PlasmoDB.org
P. berghei	genomic	http: //www.sanger.ac.uk/Projects/P_berghei/ http: //parasite.vetmed.ufl.edu/ http: //www.tigr.org/tdb/tgi/pbgi/ http: //www.GeneDB.org
P. chabaudi	genomic	http: //www.sanger.ac.uk/Projects/P_chabaudi/ http: //www.GeneDB.org
P. falciparum	genomic	http: //sequence-www.stanford.edu/group/malaria/ http: //www.tigr.org/tdb/e2k1/pfa1/ http: //www.sanger.ac.uk/Projects/P_falciparum/ http: //www.GeneDB.org
	EST	http: //fullmal.ims.u-tokyo.ac.jp/ http: //parasite.vetmed.ufl.edu/falc.htm http: //www.cbil.upenn.edu/paradbs-servlet/index.html
	GSS	http: //parasite.vetmed.ufl.edu/
	µsatellite map	http: //www.ncbi.nih.gov/projects/Malaria/Mapsmarkers/PfGMap/pfgmap.html
	optical map	http: //www.lmcg.wisc.edu/research/research.html#plasmodium
	oligonucl. array	http: //malaria.ucsf.edu/
	Affymetrix array	http: //www.scripps.edu/cb/winzeler/malariatext.html
P. knowlesi	genomic	http: //www.sanger.ac.uk/Projects/P_knowlesi/
P. reichenowi	genomic	http: //www.sanger.ac.uk/Projects/P_reichenowi/
P. vivax	genomic	http: //www.tigr.org/tdb/e2k1/pva1/intro.shtml
	GSS	http: //parasite.vetmed.ufl.edu/viva.htm
	YAC	http: //www.sanger.ac.uk/Projects/P_vivax/
P. yoelii	genomic	http: //www.tigr.org/tdb/e2k1/pya1/
	EST	http: //www.tigr.org/tdb/tgi/pygi/
Babesia bovis	EST	http: //www.sanger.ac.uk/Projects/B_bovis/
Cryptosporidium parvum	multiple	http: //CryptoDB.org
	genomic	http: //www.cbc.umn.edu/ResearchProjects/AGAC/Cp/index.htm http: //www.parvum.mic.vcu.edu/
	GSS	http: //medsfgh.ucsf.edu/id/CpDemoProj/
	EST	http: //www.ncbi.nlm.nih.gov
Eimeria tenella	genomic	http: //www.sanger.ac.uk/Projects/E_tenella/
	EST	http: //www.cbil.upenn.edu/paradbs-servlet/index.html http: //www.genome.wustl.edu/est/index.php?eimeria=1
Neospora caninum	EST	http: //www.cbil.upenn.edu/paradbs-servlet/index.html http: //www.genome.wustl.edu/est/index.php?neospora=1
Sarcocystis neurona	EST	http: //www.cbil.upenn.edu/paradbs-servlet/index.html http: //www.genome.wustl.edu/est/index.php?sarcocystis=1
Theileria annulata	genomic	http: //www.sanger.ac.uk/Projects/T_annulata/
Theileria parva	genomic	http: //www.tigr.org/tdb/e2k1/tpa1/
Toxoplasma gondii	multiple	http: //ToxoDB.org
	genomic	http: //www.tigr.org/tdb/t_gondii/
	EST	http: //www.cbil.upenn.edu/paradbs-servlet/index.html http: //www.genome.wustl.edu/est/index.php?toxoplasma=1
	BAC-end	http: //www.sanger.ac.uk/Projects/T_gondii/

Table 2. Bioinformatics resources for *Plasmodium* and other Apicomplexan parasite species

Source	Database	Tools & Features	URL
PlasmoDB	relational & flat file	query, BLAST, pattern finding, browse	http: //PlasmoDB.org
GeneDB	relational & flat file	query, BLAST, links to literature & domain databases, browse	http: //www.GeneDB.org
ToxoDB	flat file	BLAST, pattern finding	http: //ToxoDB.org
CryptoDB	flat file	BLAST, pattern finding, browse	http: //CryptoDB.org
Sanger	flat file	BLAST, software	http: //www.sanger.ac.uk
Stanford	flat file	BLAST, browse	http: //sequence-www.stanford.edu/group/malaria/
TIGR	flat file	BLAST	http: //www.tigr.org/
EMBL-EBI	flat file	query, BLAST, pattern finding, classifications	http: //www.ebi.ac.uk/parasites/parasite-genome.html http: //www.ebi.ac.uk/parasites/PlasGN/Proteome/proteome.html
TIGR Gene Indices	flat file & relational	query, BLAST	http: //www.tigr.org/tdb/tgi/protist.shtml
NCBI	flat file	BLAST, physical maps, browse	http: //www.ncbi.nlm.nih.gov/projects/Malaria/
Genome Atlases	flat file	browse	http: //www.cbs.dtu.dk/services/GenomeAtlas/
Metabolic pathways	flat file	browse	http: //sites.huji.ac.il/malaria/
Protein structure	relational	query, browse, view 3-D protein models	http: //bioinfo.icgeb.res.in/codes/model.html
SANBI	flat file	BLAST	http: //www.sanbi.ac.za/malaria-genesearch/
WEHI	relational & flat file	browse	http: //www.wehi.edu.au/MalDB-www/who.html
MR4	relational & flat file	query, browse	http: //www.malaria.mr4.org/mr4pages/index.html
PlasmoCyc	relational	enzymatic reactions, structures, list of drug targets (w/refs)	http: //plasmocyc.stanford.edu http: //www.smi.stanford.edu/projects/helix/malaria.html
UCSF	relational	query, browse transcriptome	http: //malaria.ucsf.edu

Table 3. Selected vector and host resources

Resource or Species	Contents	URL
AGRIP	database of *Anopheles* resources including mutants, methods, culture information	http: //konops.imbb.forth.gr/AnoDB/Mirror/Mbenedic/
AnoBase (formerly AnoDB)	genome database and reference center	http: //skonops.imbb.forth.gr/AnoBase/
Ensembl (Metazoan Genome Database)	genome database containing human, mouse and *Anopheles* data (among others)	http: //www.ensembl.org/
Mouse - MGI	comprehensive mouse genome and resource database (Jackson Laboratories)	http: //www.informatics.jax.org/
Mouse	multiple resources including genome, SNP, clone, gene expression,HomoloGene, LocusLink	http: //www.ncbi.nlm.nih.gov/genome/guide/mouse/
Mouse	genetic maps, physical maps, genome	http: //www.mgc.har.mrc.ac.uk/
Human	multiple resources including genome, unigene EST assemblies, physical map viewer, locus link, OMIM, SNP databases	http: //www.ncbi.nlm.nih.gov/genome/guide/human/
Human	general genome project information, Gene Gateway	http: //www.ornl.gov/TechResources/Human_Genome/home.html
Human	UK human genome mapping project resource center, with broad collection of analysis tools	http: //www.hgmp.mrc.ac.uk/
Human (Golden Path Database)	queries, BLAT, annotation for human and other genomes	http: //genome.ucsc.edu/

- Microarrays based on DNA or cDNA clones (Hayward *et al.*, 2000; Mamoun *et al.*, 2001), or oligonucleotides in either glass slide (Bozdech *et al.*, 2003a; 2003b) or photolithographic (Affymetrix) format (Le Roch *et al.*, 2002; 2003).

Protein Expression

- Tandem mass-spectrometry (Florens *et al.*, 2002; Lasonder *et al.*, 2002).

Genetic Organization and Population Structure

- Optical maps (Lai *et al.*, 1999).
- Microsatellites (Su *et al.*, 1999).
- Single-Nucleotide Polymorphisms (SNPs) (Mu *et al.*, 2002).

While it is beyond the scope of this chapter to provide a detailed description of each data type, it is worth a short digression to introduce some of the relevant technology and terms. Familiarity with the processes involved in data generation facilitates recognition of potential artifacts and sources of error, minimizing the chance of error propagation *in silico* – a risk that is inherent in computational bioinformatics.

2.1. Genome Sequence and Assembly

Two strategies are commonly employed for genome sequencing: a hierarchical approach in which the genome is broken down into smaller mapped fragments for sequencing, and a shotgun approach in which the whole-genome is subjected to random sequencing and assembly *en masse*. The former may consume considerable resources in mapping and/or fractionation of pure genomic fragments, while the latter poses greater computational problems in assembling the resulting sequence data.

The *P. falciparum* genome was sequenced using a hierarchical approach, in which chromosomes were separated in pulse field gels (and some chromosomes were further sub-cloned into YACs) prior to the production of random clone libraries for sequencing (Gardner *et al.*, 1998; Bowman *et al.*, 1999; Gardner *et al.*, 2002a; 2002b; Hall *et al.*, 2002; Hyman *et al.*, 2002). The *P. yoelii* genome was sequenced using a whole-genome shotgun approach (Carlton *et al.*, 2002).

In either approach, sequences generated by random sequencing (of either the entire genome, or individual chromosomes or smaller fragments) must be reassembled into larger pieces of contiguous DNA, or "contigs", and several software packages are available for contiguating sequence reads. Contigs are then organized into larger "scaffolds" containing gaps between the individual contigs, based on information from mapping data, end sequences from large-insert clones, etc. Overall, genome assembly is a tricky business; common problems and sequence artifacts include:

- Repetitive regions of the genome can cause mis-assembly errors, i.e. sequences that are not adjacent to one another in the genome can become artificially merged if they contain identical stretches of sequence (repeats). While the *P. falciparum* genome is relatively small (by eukaryotic standards), and 'satellite' DNA and other repetitive sequences are not particularly abundant, the extremely high A+T nucleotide content raises similar problems: unique sequences are hard to find in a two letter alphabet!

- When hierarchical sequencing approaches are used, DNA sequenced from one fraction (e.g. one chromosome) may be contaminated with DNA from another fraction (chromosome). Thus, genes may initially appear to be located on the wrong chromosome, or on multiple different chromosomes. Such sequences usually do not contiguate with the

majority of the sequences and appear as "singlets" or orphan sequences until sequences for the entire genome are pooled and compared for final assembly. Early stages of assembly of the *P. falciparum* genome contained many overlapping fragments of HRP2, a known single-copy gene. Once the entire genome is assembled, most the remaining orphan sequences are likely to be attributable to contaminating DNA, mis-assembled sequence reads, and other artifacts ... but these sequences will undoubtedly include some valid sequences as well, including RNA- and protein-coding genes.

- Shotgun sequencing provides an extremely cost-efficient means to identify most sequences, but the laws of probability and combinatorics ensure that some sequences will be missed. Moreover, because genomic sequences differ in their clonability, not all will be represented in the library (a problem that is particularly acute for A+T-rich genomes), shotgun sequencing rarely achieved the theoretical level of sequence coverage. Thus, while the 5X random shotgun sequence available for *P. yoelii* means that 5 genome equivalents of DNA have been sequenced (>100 Mb), the assembled sequence still contains many gaps (>5000). Closing such gaps is a laborious and expensive process, and as of this date the *P. falciparum* genome still contains a few physical gaps and a few "unmapped" regions of sequence that need to be correctly placed in the genome.

- If the sequence reassembles into multiple pieces, how should these pieces be ordered and oriented? In the case of the *P. falciparum* genome, two sets of physical anchors or genome landmarks were available to help order the fragments along the chromosomes into scaffolds: a microsatellite map (Su *et al.*, 1999) and an optical map (Lai *et al.*, 1999). *P. falciparum* chromosomes 6-8 (affectionately known as the "BLOB") could not be resolved on a pulse-field gel, and posed a particular challenge. Additional "Happy maps" were therefore constructed to facilitate the ordering and assembly of these chromosomes (Hall *et al.*, 2002).

- Contaminating sequences from cloning vectors (plasmids, transposons), cloning hosts (*E. coli*, yeast), human DNA and other organisms being sequenced may enter into raw sequence output. Users of sequence data should investigate any suspected cases of horizontal gene transfer very carefully, especially at the DNA nucleotide level and via genomic Southern blots, to guard against such sequencing artifacts.

Each of the above difficulties can be resolved, but users examining pre-publication data are cautioned to be aware of potential sequence artifacts. *P. falciparum* sequences have been cleaned of most artifacts, but sequences for other species still have problems. For example, the *P. reichenowi* sequence is known to be heavily contaminated with monkey DNA.

2.2. Maps

Genetic and physical maps consist of markers at specific genetic or physical locations within the genome. Classically, these have been constructed based on cytogenetic banding patterns, or by using genetic crosses to map loci responsible for various phenotypes. Unfortunately, because the nuclear envelope does not break down during mitosis in *Plasmodium* (as in most protozoa), it is not possible to isolate condensed chromosomes for the analysis of banding patterns. Genetic mapping studies based on classical genetic crosses are feasible, however, and 1 cM in *P. falciparum* has been measured as ~17 kb (Su *et al.*, 1999). (Centimorgans provide a standard measure of recombination frequency; 1 cM represents an average distance of 10 Mb in humans). Due to the difficulty of conducting classical genetic crosses and mapping of phenotypes in *Plasmodium,* however, alternative approaches have also been developed. Both microsatellite and optical maps of the *Plasmodium* genome have facilitated ordering of the hundreds of genomic contigs onto chromosome scaffolds used for genome closure.

Optical maps rely on novel imaging technology to construct a chromosome-scale restriction enzyme map. Large fragments of chromosomal DNA are attached to a solid surface under gentle fluid flow. After adhesion, a restriction enzyme is added to cleave the DNA *in situ,* leaving an ordered line of fragments, whose size can be assessed microscopically based on labeling with an intercalating DNA dye (providing a quantitative measure proportional to DNA content). Optical maps have been created for *Plasmodium falciparum* strain 3D7 using two different restriction enzymes (Lai *et al.*, 1999).

Microsatellite maps are based on the use of PCR primer pairs that amplify regions differing in length between the two parental genomes, converting a sequence (length) polymorphism into a genetic marker (Su *et al.*, 1999). By examining the lengths of the PCR products in the progeny of a genetic cross, it can be determined which regions of the genome came from which parent. Restriction polymorphisms can also be employed for genetic mapping studies, although these have proved more cumbersome in *Plasmodium,* in part because of the high A+T content of the genome. Careful association of microsatellite patterns with phenotypic or genetic markers permits the construction of an integrated genetic and physical map, linking individual microsatellites to particular regions of specific chromosomes. In addition to their utility for genetic mapping, including the analysis of Quantitative Trail Loci (QTL) for specific phenotypes (Ferdig and Su, 2000; Wootton *et al.*, 2002), microsatellite markers are also extremely useful tools for population surveys.

2.3. Expressed Sequence Tags

Expressed Sequence Tags (ESTs) are sequences obtained from reverse-transcribed mRNAs (cDNAs). As such, they can be used to determine gene structure (exons and introns), and identify open reading frames that are likely to encode protein sequences. By focusing on transcribed sequences and minimizing the problems associated with splice-site prediction, EST projects are extremely cost-efficient, delivering a large number of protein predictions for relatively small cost. Moreover, the abundance of EST sequences obtained for individual genes provides a crude indication of transcript abundance in the original library. EST sequences are currently available from several *Plasmodium* cDNA libraries, representing various life cycle stages and species. While the collection and analysis of such data has been discussed elsewhere (Ajioka *et al.*, 1998; Li *et al.*, 2003), several issues likely to impact on bioinformatics experiments are worth considering here.

- Because EST sequences are derived from specific libraries, they represent only the individual strains and life-cycle stages from which these libraries were generated, and the transcripts produced by those parasites. Thus, while random sequencing of genomic DNA can in theory approach complete representation of the parasite genome, no EST project is likely to provide a complete catalog of all genes. Representation is sometimes enhanced, however, by using normalized libraries in which highly abundant sequences have been suppressed (using a variety of strategies).

- EST abundance may be able to provide a crude estimate of relative transcript abundance, but such estimates are likely to be biased by library amplification, and completely invalid in normalized libraries (depending on the method employed). In general, hybridization with RNA, RT-PCR, SAGE, and microarray analysis (see below) provide more suitable methods for transcript profiling.

- EST sequences are often incomplete. To facilitate gene discovery, most EST projects use libraries of directionally-cloned cDNAs (although note that up to 30% of inserts may be cloned in the inverted orientation), and produce only a single sequence read from the presumed 5′ end of the cDNA. Because many cDNA clones do not represent full-length mRNAs, however, the start of the sequence may not provide the transcript initiation site, especially for long mRNAs. A single EST sequencing reaction typically yields ~350 nt, and is therefore very unlikely to provide the complete cDNA sequence. Repeated sampling of the library often yields multiple overlapping ESTs derived from the same gene, and clustering of these may yield a longer consensus sequence (see below), but most of these sequences still remain incomplete.

- EST sequences are likely to contain a higher error rate than genomic sequences, for a variety of reasons. For example, because EST projects focus on gene discovery rather than high-fidelity genome assembly, individual cDNAs are generally sequenced only once, although transcripts derived from the same gene may be sequenced multiple times in a given library, as noted above. Clustering of ESTs can help to extend sequence length and reduce error, but inclusion of data from multiple isolates may yield an inaccurate consensus whenever allelic polymorphisms are present. (Indeed, correlation of multiple sequence alignments with strain information provides an excellent source of microsatellite and SNP markers for genetic analysis.) It is therefore important to distinguish between individual ESTs and consensus sequences, and to recognize that even sequences derived from the consensus of many ESTs may be incorrect.

- Differentially-spliced genes – while far less common than in metazoan species – are nevertheless well known in *Plasmodium,* and it is important to recognize when differentially-spliced transcripts derive from the same gene, as opposed to paralogous genes or strain-specific allelic variants. Note, however, that incomplete intron excision is quite common, producing many cDNAs that are unlikely to be fully translated.

- As with genomic sequences, EST libraries may contain a low frequency of contaminating sequences. The lack of redundant sequencing makes it difficult to distinguish rare transcripts from contaminating DNA, however; putative transcripts should always be validated by comparison with genome sequence (when available) and hybridization with genomic DNA. cDNA libraries may also be contaminated with incompletely processed sequences, and with genomic DNA, yielding apparent transcripts that are unlikely to be translated and may not even be transcribed. This is particularly problematic for *Plasmodium,* where the high A+T content may lead to false priming by oligo-dT. In addition, an early Genome Survey Sequence (GSS) project using mung-bean nuclease libraries to provide a 'genes-first' approach to sequencing of *P. falciparum* DNA produced sequences that were initially mislabeled as ESTs.

- Because EST sequencing is often a continuing project, the identifiers associated with assembled sequences for an individual gene may change frequently, producing considerable confusion. In practical terms, it is often most convenient to find the new consensus sequence (and name) via a BLAST query with the old sequence. It is also critical to note the data release date and database version on which any analysis is based.

When ESTs are analyzed in bioinformatics experiments, it is often necessary to understand how the data were generated in order to interpret the results correctly. What species, strain, stage? Is the library directional?

amplified? normalized? How were contaminating sequences removed? How were individual ESTs assembled? How many sequences are represented in an individual cDNA assembly, how well are they aligned, and how deep is the alignment as a function of position?

2.4. SAGE Tags

SAGE (Serial Analysis of Gene Expression) was developed to provide a "snapshot" of mRNA abundance at a given time in a given cell or tissue type (Velculescu *et al.*, 1995). Rather than conducting a full-scale EST project, sequencing cDNAs in their entirety, each EST is reduced to a short oligonucleotide sequence tag, normally near the 3′ end of the gene. These tags are then ligated into large concatemers, so that an individual sequencing reaction can identify tags derived from dozens of individual cDNAs, rather than the single cDNA represented by an individual EST sequence. Computational analysis is used to determine which gene/mRNA matches which specific SAGE tag. Once again, the high A+T content of the *Plasmodium* genome poses a problem. For example, the GC-rich 10mer SAGE tag GGTTCAGGGT is predicted to occur 0.59 times by chance in the *P. falciparum* genome (based on the observed 80% frequency of A+T), while the AT-rich tag ATCATATAAG is predicted to occur 150 times by chance alone. Thus, the mapping of SAGE tags and other short oligonucleotides to the *P. falciparum* genome may be a "one-to-one" or a "one to many" relationship, but when SAGE tags are combined with other data (gene predictions, EST sequences, BLAST similarities, etc) it is often possible to determine the true sites of expression (Munasinghe *et al.*, 2001; Pleasance *et al.*, 2003) .

2.5. Transcript Expression Profiling

Several types of microarrays may be employed to examine the expression of many individual genes in parallel. All of these methods involve immobilized gene-specific nucleic acids: genomic DNA clones, cDNA clones (ESTs), or synthetic oligonucleotides. Plasmid clones can easily be isolated from genomic libraries, and libraries constructed using mung-bean nuclease (which cleaves preferentially in extremely AT-rich DNA) may favor clones containing individual genes or gene exons. cDNA clones can be isolated from plasmid libraries as well, and offer the advantage of being unequivocally derived from individual mRNAs (subject to the quality of the library), although they are unlikely to represent the genome as a whole – with highly-expressed genes represented many times, and other genes not represented at all. All clone-based reagent sets are problematic from the standpoint of quality control: reliably propagating, quantitating, and tracking thousands of individual plasmids is a daunting task. As a result, most array

projects have now moved to oligonucleotide-based microarrays, provided that effectively complete genome sequence is available – as is indeed the case for *P. falciparum* (and many related species).

Two alternative formats for oligonucleotide-based microarrays are in common use. In the first, oligonucleotides are synthesized in 96- or 384-well format, and robotically spotted onto glass slides. This format offers several advantages, including the ability to design custom microarrays tailored to individual experimental needs, and the ability to design new probes to take advantage of improved genome annotation and new experimental approaches. Disadvantages include the cost of oligonucleotide synthesis (or purchase), the potential for error in reagent generation/storage/tracking, and difficulties in maintaining the spotting robots for reproducible array production (although many large research centers now support microarray facilities).

Alternatively, oligonucleotides can be synthesized directly on the microarray, using proprietary photolithographic methodology. This format offers the ability to print features at higher densities (typically 500,000 features/chip, vs ~15,000 for glass slide arrays), with greater reproducibility. Disadvantages include the proprietary nature of the technology involved, the cost and inflexibility of photolithographic array design, and the high cost of arrays (obtainable only through the Affymetrix Corporation), and the need for sufficiently large orders to justify printing. Facilities for reading microarrays in both formats are generally available at most large research centers.

For *P. falciparum,* glass slide microarrays have been produced containing cDNA sequences (Mamoun *et al.*, 2001), spotted mungbean nuclease fragments (Hayward *et al.*, 2000), and 70mer oligonucleotides representing the vast majority of predicted genes in the genome (Bozdech *et al.*, 2003a; 2003b). Oligonucleotide probe sets for *P. falciparum* are now available commercially (www.qiagen.com/arrays/oligosets_malaria.php). An Affymetrix chip containing shorter oligonucleotides for nearly every predicted exon in the *P. falciparum* genome, as well and non-coding and opposite strand regions (at a lower frequency) has also been designed (Le Roch *et al.*, 2002; 2003). Many of the resulting expression data sets have been deposited in the *Plasmodium* Genome Database (http://PlasmoDB.org), and the sequences used to create these arrays or oligos have been mapped to the *P. falciparum* genome. For example, the expression profile for the major merozoite surface protein (MSP1) of *P. falciparum* is shown in Figure 1.

An in depth discussion of all bioinformatics aspects of microarrays is beyond the scope of this chapter, but as with all studies producing genomics-scale datasets, it is critical that experiments be well-controlled, reproducible, and understood by any user hoping to make sense of this data, particularly as data from different types of experiments are often stored and accessed in different ways. For example, in comparing two whole-genome expression

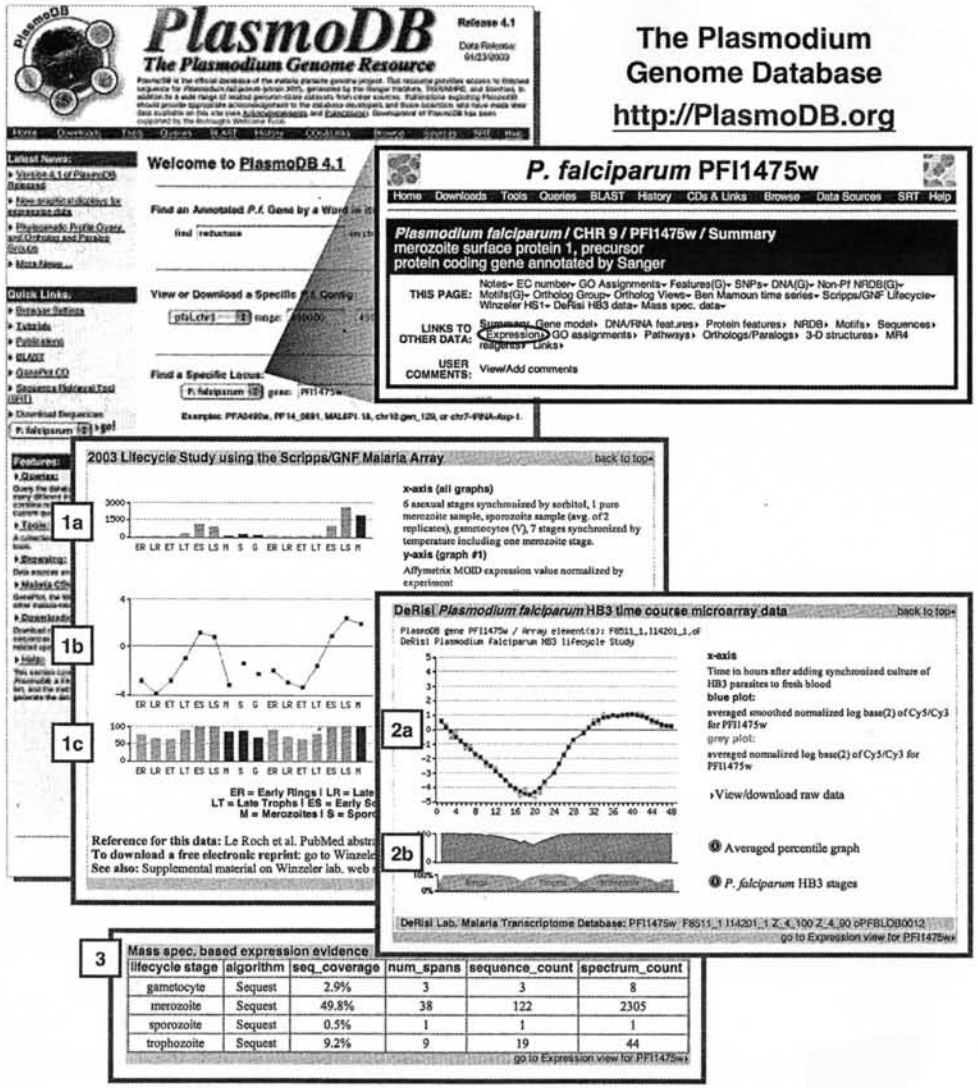

Figure 1. Browsing *P. falciparum* genes in PlasmoDB 4.1. Multiple alternative views are available for individual genes. The "expression" view of MSP1 (merozoite surface protein 1; PFI1475w) presents information on RNA and protein expression, including data from both Affymetrix and glass-slide microarrays, and proteomics analysis by MS/MS. See text for further details.

profiling datasets for *P. falciparum* that are accessible via PlasmoDB, the "Scripps/GNF" Affymetrix array (Le Roch *et al.*, 2002; 2003) provides absolute expression values (Figure 1, graph 1a) for seven time points spanning the intraerythrocytic life cycle (using two independent synchronization methods), in addition to data on expression in sporozoites and gametocytes. The spotted glass slide arrays reported in (Bozdech *et al.*, 2003a; 2003b) use a different parasite strain (HB3 vs 3D7), and provide higher time resolution: 48 hourly time points across the intraerythrocytic life cycle. Because experimental variability is high for glass slide arrays, expression values are normalized to a common pooled control, yielding a graph of expression induction rather than absolute values (Figure 1, graph 2a). In order to enable direct comparison between these two microarray platforms, the Scripps/GNF dataset is also presented in the form of of induction ratios (Figure 1, graph 1b), and absolute expression levels are presented as a percentile of all genes in each experiment (Figure 1, graphs 1c and 2b). Both experiments indicate high abundance and strong up-regulation of steady-state MSP1 transcript levels in late schizonts. Raw data can also be downloaded, and links are provided to the home sites for all relevant data sources. All probes are mapped to the parasite genome, enabling convenient comparison.

2.6. Proteomics

The production of large-scale proteomic datasets has been made possible by technological advances in mass spectrometry, combined with the availability of complete genome sequences. Analysis of complex protein mixtures (as opposed to purified proteins) and the determination of putative peptide sequences (as opposed to the masses of proteins or peptide fragments), permits comparison with predicted sequences emerging from genome sequencing projects (although the scale of whole genome computational analysis can be problematic). To date, two major genomic-scale proteomic analyses have been published for *P. falciparum* (Florens *et al.*, 2002; Lasonder *et al.*, 2002). Such studies provide a snapshot of the protein repertoire at a given time, and have permitted recognition of >40% of all annotated proteins in the parasite genome.

As with the analysis of microarray data, it is imperative to understand the nature of the data obtained, and limitations of the available results. Because peptide recognition depends on gene predictions, protein sequences associated with incorrectly assigned gene models will not be recognized. Searches of all open reading frames in the genome may lead to the discovery of a gene that was expressed but not predicted in the genome sequence. Even when gene models are accurate, many factors may influence the ability to detect peptide sequences, including protein abundance; post-translational modification; efficiency of solubilization; proteolytic digestion, and ionization; etc. Thus, positive data is likely to indicate peptide presence, but

negative data is far less informative. Note also that analyses conducted to date provide no reliable quantitation of abundance, although the number of peptides recognized may provide a crude indication: MSP1 was identified in all samples, but was represented by far more peptides, covering far more of the gene, in merozoites (Figure 1, graph 3).

3. Data Repositories and Organization

The *Plasmodium* Genome Database, PlasmoDB (http://PlasmoDB.org), provides the largest and most comprehensive single collection of *Plasmodium*-related data (Bahl *et al.*, 2002; Kissinger *et al.*, 2002). This community resource currently houses genome sequence for several *Plasmodium* species; multiple alternative gene predictions; automated and curated annotation, including controlled vocabulary Gene Ontologies; GSS and EST sequences; SAGE data; mungbean, cDNA and oligonucleotide microarray data based on both glass slide and Affymetrix platforms; MS/MS proteomic data; microsatellite and physical mapping data; and comparative genomic analyses. Much of the data available in PlasmoDB is also available on CD-ROM. Depending on the application, the reader will also benefit from various other *Plasmodium* and apicomplexan parasite resources, as discussed below. Table 1 summarizes sites where data are stored and often available for download. Table 2 lists sites supporting bioinformatics analysis and data queries.

Each of the sequencing centers involved in the generation of the *P. falciparum* genome (the Sanger Institute, Stanford University, and The Institute for Genome Research; TIGR) maintains a BLAST searchable website and an FTP download site where sequences generated by that center may be obtained. Gene predictions and features may be queried at the Sanger Institute via GeneDB. TIGR maintains an EST-based gene index (Quackenbush *et al.*, 2001) for *P. falciparum* and *P. yoelii* (as well as several other apicomplexan parasites), offering a non-redundant view of transcripts analyzed computationally to provide information on potential cellular roles and function. In order to illustrate the logical sequence of events required for developing a bioinformatics resource, Box 1 summarizes the strategy for Gene Index production.

Several databases are dedicated to metabolic pathways and drug discovery. The "Malaria Parasite Metabolic Pathways" site (Table 2) provides curated graphical snapshots of *Plasmodium* metabolic processes organized by pathway. Direct links are provided from each enzyme E.C. number to Expasy-NiceZyme views, Brenda (Schomberg *et al.*, 2002) and PlasmoDB databases. PlasmoCyc contains graphical and searchable representations of *P. falciparum* metabolic pathways, and a whole-cell overview of metabolic pathways along with tools for between-species comparisons. The resource

Box 1. The TIGR Gene Index Protocol for Assembly of ESTs and Transcripts (http: //www.tigr.org/tdb/tgi/definitions.html)

Preparation of EST data

- Extract sequences from dbEST and subject to quality control screening (vector, *E. coli*, polyA, T, or CT removal, minimum length = 100 bp, < 3% N).

Preparation of transcript (ET) database

- Extract all sequences from the appropriate division of GenBank.
- Discard non-coding sequences.
- Save cDNAs and coding sequences from genomic entries.
- Store sequences and related information in Expressed Gene Anatomy Database (EGAD).
- Make curated ET data set available as a multiple FastA format file (see EGAD main page).

Assembly

- Combine cleaned EST sequences and non-redundant transcript (ET) sequences.
- Assemble sequences into contigs using Paracel Transcript Assembler Program. TCs are consensus sequences based on two or more ESTs (and possibly an ET) that overlap ≥40 bases with ≥94% sequence identity (strict criteria help minimize creation of chimeric contigs).
- Assign contigs a TC (Tentative Consensus) number. TCs may comprise ESTs derived from different tissues.
- Assign best hits for TCs by searching against a non-redundant amino acid database (nrAA) using BLAT.
- Select and display top five hits (based on score) for each TC.

Caveats

- TCs are only as good as the underlying ESTs; unspliced or chimeric ESTs will produce aberrant TCs.
- The TC set contains some redundancy because sequences will not be combined unless they exhibit a high % identity and match end-to-end.
- TS directionality should not be assumed.
- Not all TCs contain protein-coding regions.

can also display individual enzymatic reactions with substrate and reactant structures, cellular localization, information regarding the association of protein subunits into complexes and a list of predicted drug targets (with links to the papers describing them).

Protein annotations have been used to search the protein structure database (PDB), and several *P. falciparum* protein structures have been modeled and are viewable. Microarray data are available from several sources (Table 1) and the UCSF site provides extensive viewing and analysis capabilities (Table 2). DNA structural analyses (repeat content, DNA "bendabilty", etc.) have been calculated for *P. falciparum* and can be viewed using the genome atlases maintained at the Center for Biological Sequence Analysis (CBS). The WHO/TDR Malaria database contains searchable genome annotation and an electronic repository of *Plasmodium* strain information, antigen

and other multiple sequence alignments, and an extensive malaria antigen literature database; this database is also available on CD-ROM.

The Malaria Research and Reference Reagent Resource Center, MR4 (Adams *et al.*, 2000) is both an electronic and physical repository for quality controlled malaria-related reagents and information. Registered users can obtain parasites, mosquito vectors, antibodies, antigens, clones and gene libraries. MR4 resources are searchable via the web and many reagents and/or genes have been linked to PlasmoDB (and vice-versa).

Dozens of other databases are extremely useful for *Plasmodium* research. While it is impossible to describe all of these sites here, Tables 1-3 provide a compendium of several such resources, and most are described in detail in the annual database issue of Nucleic Acids Research (Baxevanis, 2003). Useful databases include (but are not limited to): the NCBI GenBank and EMBL databases, containing large sequence repositories and a variety of tools for accessing these data; *Anopheles*, mouse and human genome databases (Table 3); the SMART (Simple Modular Architecture Research Tool) database for examining protein domain architectures; the BIND (Biomolecular Interaction Network Database) of molecular interactions culled from the literature and high-throughput analyses; the PFAM (Protein Families) database, containing a collection of Hidden Markov Models (HMMs) used to screen protein sequences and identify protein family members based on conserved patterns; and the ProDom and InterPro protein domain databases for searching and identifying protein domains.

The data described above is stored in a variety of formats. "Flat file" text documents can be opened in a word processor or spreadsheet program, and are therefore easy to share. Search functions are generally limited to "Find" commands to locate key words, and/or "Sort" commands to arrange and/or manipulate data. Even very large datasets, such as BLAST databases are typically stored in flat-file format. Other database types – relational and object-oriented – permit more sophisticated functions, and are usually managed by a database management system (DBMS) such as Oracle, DB2 to keep track of data access, deposition, security etc. Such management systems keep track of data records and requests for their access, preventing (for example) simultaneous withdrawal of checking account funds in excess of the amount on deposit.

3.1. Relational Databases and Queries

Relational databases store data in tables that are designed to accommodate specific data types, as shown in Figure 2. For example, one table might contain the names of all students in a school, another table might contain the names of all professors, a third the names of all the classes offered, and a fourth, the list of rooms in which classes are taught. Each of these tables

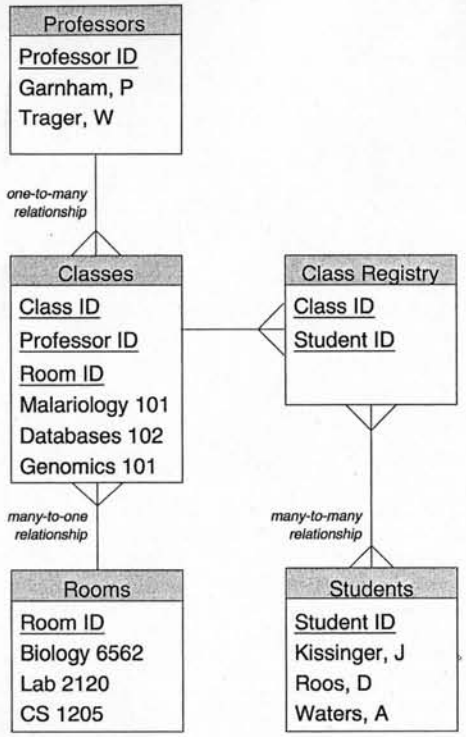

Figure 2. A relational database schema. Each data type (Professors, Classes, etc) is stored in a distinct table. Conceptually, tables contain both columns and rows. Tables can hold multiple entries for each data type: a unique identifier ID, the data itself (e.g. names of professors), and identifiers linking data in one table to an ID in another. Tables are associated with each other according to the type of relationship (one-to-one, one-to-many, many-to-one, many-to-many). In this example, each class is taught by a single professor, but each professor may teach multiple classes. Classes have multiple students, and students have multiple classes. An additional Class Registry table links each student to each class. Each class is associated with a single classroom, but an individual classroom may be used for multiple classes. Professors are not directly linked to rooms or students, but a relationship is defined via the classes that they teach.

can accept certain values (e.g. names consisting of alphabetical characters up to 50 characters in length, or alpha-numeric codes like Biology 6562 for classroom location). Relational databases, in addition to storing the data in defined tables, also relate the data contained in the tables to one another. For example, in the example shown, professors teach specific classes, classes have students enrolled in them, and classes are taught in specific classrooms. These relationships are not random, but clearly specified: each class has only one classroom, only one instructor, but multiple students. Students enroll in classes, and classes are taught in classrooms, but there is no direct link between students and classrooms; these two tables are only related via the classes offered. If the appropriate relationships are specified in the design of

the database, then users of the database can ask questions (called queries) that require data from multiple tables. For example, one might want to know all classes taught by professor P.C.C. Garnham, all classes taught in the Biology 6562 classroom, or the names of all students enrolled in Malariology 101.

Extending this analogy to consider the very diverse kinds of biological data relevant to *Plasmodium* parasites, database tables can be generated to hold genomic sequences, EST sequences, SAGE tags, translated protein sequences, protein domains, peptides identified by MS/MS, and microarray oligonucleotides, etc. Relationships can then be defined between these tables: protein domains and peptides determined by MS/MS may relate to particular EST sequences or gene models; oligonucleotides from expression studies can be related to predicted or annotated genes; these genes can be related to proteins and these proteins can be related to function via Gene Ontology classifications; etc. The PlasmoDB database is built on a relational schema (Genomics Unified Schema; GUS) that currently contains more than 200 tables (Davidson *et al.*, 2001).

3.2. Controlled Vocabularies

Meaningful comments or data analysis requires controlled vocabularies – a standard set of terms that are applied to equivalent genes or processes. For example: it is often necessary to search for a gene by name, but what is that name? MSP-1, Merozoite Surface Protein-1 and PFI1475w are synonymous. In order to create automated systems for comparison it is necessary to agree upon a common vocabulary that is used for all organisms, such as the enzyme commission (E.C.) classification system.

Gene Ontology (GO) terms provide another example of controlled vocabularies. GO terms are created and maintained by the "Gene Ontology" Consortium (http://www.geneontology.org), as hierarchies of increasingly generalized terms around the concepts of "Molecular Function", "Biological Process", and "Cellular Component" (Ashburner *et al.*, 2000). By design, these definitions are sufficiently flexible that they can evolve as new information becomes available.

For example, GO terms for MSP-1 include:

- Biological process:
  - GO:007154, cell communication
  - GO:0030260, cell invasion
- Cellular component:
  - GO:0005623, cell
  - GO:0016020, membrane
- Functional assignment: Not yet defined

Each of these assignments was made by an annotator and comes with an evidence code describing the basis used when assigning the term (in this case all are labeled "TAS", or traceable author statement). Other GO evidence codes include: IC, inferred by curator; IDA, inferred from direct assay; IEA, inferred from electronic annotation; IEP, inferred from expression pattern; IGI, Inferred from genetic interaction; IMP, inferred from mutant phenotype; IPI, inferred from physical interaction; ISS, inferred from sequence or structural similarity; NAS, non-traceable author statement; ND, no biological data available; and TAS, traceable author statement. See http://www.geneontology.org/doc/GO.evidence.html for full explanation.

Once GO terms are applied to gene products, many of the problems related to data searching, integration and comparisons become much simpler. Searches can be performed using GO terms, and genes can easily be related – even across species boundaries –using GO identifiers. Evidence codes provide users with a clear statement as to the origin and confidence associated with each assignment. The combined information provided by GO term classifications and their evidence codes informs database users of what is known about any given assignment.

3.3. Data Integration

Data integration is the process of relating one type of data to another. As a simple example, gene annotations are related to a particular genome sequence, and protein features to protein sequences. Of course, integration can also involve more diverse data types: sequences may be related to physical maps, microarray oligonucleotide probes, or proteomic fragments. Defining data integration linkages is a laborious task, requiring extreme attention to detail. One common type of data integration performed by researchers on a regular basis is the association of gene names (and hence putative function) with a given sequence. Such relationships are often inferred on the basis of similarity to other sequences or the presence of particular motifs, in which case, a similarity search (e.g. BLAST) may provide the information necessary to link two diverse pieces of data. Sequence similarity searches are also commonly used to relate EST sequences to genomic sequences.

Applying such processes to genomic-scale datasets can yield very large networks of integrated data. Since nearly all data types can be related to the genome sequence either directly or indirectly, the genome sequence becomes a "bridge" that allows diverse data types to be integrated. Recall the database example provided in Figure 2. While professors are not directly related to classrooms, these distinct data types are related and integrated via the classes offered. Applying the same reasoning to genomic data, it is possible to link a proteomic mass profile to microarray expression levels, along the following path: collision-induced ionization of peptide fragments produces masses,

the difference between these masses corresponds to a particular amino acid sequence, this sequence can be found in the data set of predicted proteins, predicted proteins are related to open reading frames, mRNAs are related via gene predictions to regions of genomic sequence, portions of the genomic sequence may correspond to cDNAs or oligonucleotides on a microarray, and elements on this microarray are linked to transcript expression levels, completing the path.

Controlled vocabularies have been developed for a variety of data types, and greatly facilitate the establishment of relationships and integration of diverse data types. With the advent of GO terms and motif identifier numbers and names, these data can be used to quickly relate homologous genes across species, and to identify all predicted proteins with a particular domain arrangement. Further integrating ProDom motif terms with GO identifiers may be able to ease the laborious task of assigning GO terms (Schug *et al.*, 2002): if protein motifs in a predicted sequence can be associated with ProDom terms, and a particular motif order and/or combination can be associated with a GO terms, then automated integration could facilitate annotation. Other data types require unique solutions, such as establishing the locations of all potential SAGE tag origination sites, or the location of enzyme restriction sites that would give rise to the fragments (± error) observed in the optical mapping experiments. Making the data from such analyses accessible through a relational database allows the full power of the integration to be realized.

Great care is essential in defining database schema architecture, and loading data into the database, however, as errors in data integration (such as applying gene prediction coordinates from one genome assembly to another) can be disastrous. Such challenges are a particular concern in a highly networked database architecture where the underlying data is constantly changing. When the reference genome sequence changes, for example, all data must be re-integrated. To minimize potential problems, it is important to record version numbers for each and every data source utilized. This subject is discussed more fully below.

3.4. Working in a Mixed Database World

The term "database" can be applied to many forms of collected data that can be downloaded, browsed, analyzed and/or queried. Tables 1-3 list many of the web-accessible data sources and analysis tools available for Apicomplexa and some of their host and vector species. In general, web sites incorporating search tools where user-specified text (or items from pull-down menus) can be used as search terms are likely to be based on a relational architecture. Such databases may indicate they were built using Oracle, Sybase, mySQL or PostgreSQL. Any database that indicates queries can be constructed using a Structured Query Language (SQL) is relational.

It is often impossible to identify the functionality and/or data types available at any given site without exploring the site and reading the introductory and/or tutorial pages, as database "look and feel" is more an indication of artistic style than functionality. The appearance of a web-based "front-end" may be held constant, even when the underlying database architecture is changed. For example, many aspects of PlasmoDB initially handled by smaller flat-file databases have become incorporated into the GUS relational schema. Conversely, changes in the appearance of a web-based "front-end" need not reflect any change in the underlying architecture. The appearance of the NCBI GenBank and PlasmoDB have both changed over the years, while maintaining similar core services and taking on additional functionalities; a little exploration reveals how new tools have been implemented. Similarly, quite different interfaces may be used to access similar databases. For example, the GUS architecture employed for PlasmoDB has recently been adopted by the Sanger Institute to drive GeneDB, providing access to the various organisms sequenced and annotated by Sanger's Pathogen Sequencing Unit. This development should greatly facilitate the development and exchange of software for browsing, visualizing, analyzing, mining and querying data.

Because genomic datasets change frequently, databases – like the web pages used to access them – typically display version numbers or release dates, which should be noted in any publications that depend on these resources, and any communications aimed at identifying problems in data access, analysis, or integration. It is not uncommon to discover that different identifiers are used for identical data stored in different databases, or that the same identifiers may be used for different data in different databases, or different releases of the same database. With the ever-expanding availability of internet resources, the possibilities for confusion are endless! It is therefore critical that bioinformatics researchers keep track of information on data release dates, database versions, etc, and make this information available in any publications (print or electronic) that may result. It is equally important that database developers provide resources that allow published data to be tracked and updated (enabling the correlation of new and old EST assemblies, for example), or at least maintain the ability to access old release data.

Particular attention should be paid to the specific data sets and analysis tools provided in the various available databases. For example, NCBI BLAST and Washington University BLAST (WU-BLAST) are different implementations of the same local sequence alignment algorithm; both work well, but they employ different default settings, arguments and DNA scoring matrices, and will therefore yield slightly different results. Most large genome databases provide a combination of human curated and computationally generated automated analyses. The availability of such diverse data types is what makes bioinformatics analysis possible, but it is important to know exactly how the data have been curated and/or analyzed (see Box 1 for an example).

4. Queries: Powerful Tools for Developing Bioinformatics Prowess

Although most databases permit the data that they house to be examined (for example, scrolling through all genes on *P. falciparum* chromosome 1), and many provide tools for data analysis (e.g. BLAST searches against *P. vivax* sequences, or listing all abundant transcripts in gametocytes in a particular experiment), the real power of a relational database lies in the ability to form integrated queries that depend on the relationships between multiple data types, as defined in the database schema. A wide range of queries are available in PlasmoDB, including queries related to

- curated or automated annotations (including GO function, process and component)
- chromosomal location
- results from BLAST searches against GenBank and/or other *Plasmodium* species
- DNA sequence features: low complexity sequence, AT content, coding potential, etc
- gene structure (intron-exon architecture)
- protein sequence features: secretion and organellar targeting signals, transmembrane domains, Pfam/ProDom/other motifs, secondary structure, predicted CD8 epitopes, etc
- the presence of strain-specific nucleotide and/or amino acid sequence polymorphisms
- expression data: EST or SAGE abundance, RNA expression levels/ induction/timing (on several platforms), evidence for protein expression (from MS/MS analysis)
- phylogenetic cross-comparisons (in comparison with other *Plasmodium* species, other Apicomplexa, other eukaryotic species, etc)
- availability of a predicted protein structural model
- involvement in specific metabolic pathways
- availability of reagents

In developing a successful query, it is crucial to translate biological knowledge into computationally-accessible terms. For example, a query for "drug targets" is not particularly well-defined, but a query for enzymes for which a structural model is available and that are known to be expressed in the erythrocytic stages at both the RNA and protein level, can take advantage of GO terminology, the *Plasmodium* protein structural model database, and both microarray and proteomics datasets (Kissinger *et al.*, 2002).

To pursue a vaccine-related example, one might wish to look for surface antigens based on the presence of a predicted secretory signal sequence (and/or one or more transmembrane domains), as shown in Figure 3. In addition, one could look for antigens shared between *P. falciparum* and

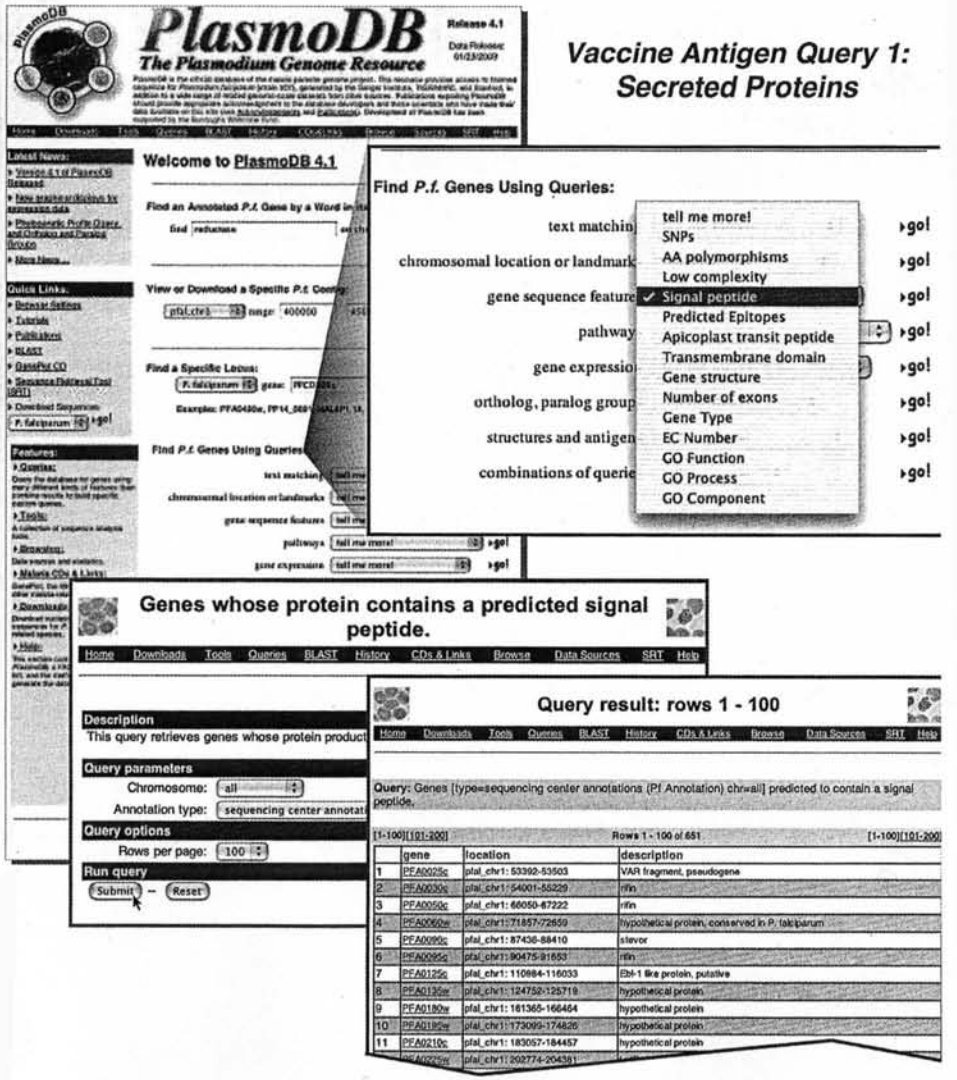

Figure 3. Querying the curated annotation of *P. falciparum* genes in PlasmoDB 4.1 for signal peptides (identified using the neural net work program SignalP) yields 651 proteins that are predicted to be secreted. Because accurate prediction of secretory signal sequences requires accurate assignment of the translational initiation, it is likely that this query misses many secreted proteins. Further refinements might include searching alternative gene models, or including proteins with predicted transmembrane domains.

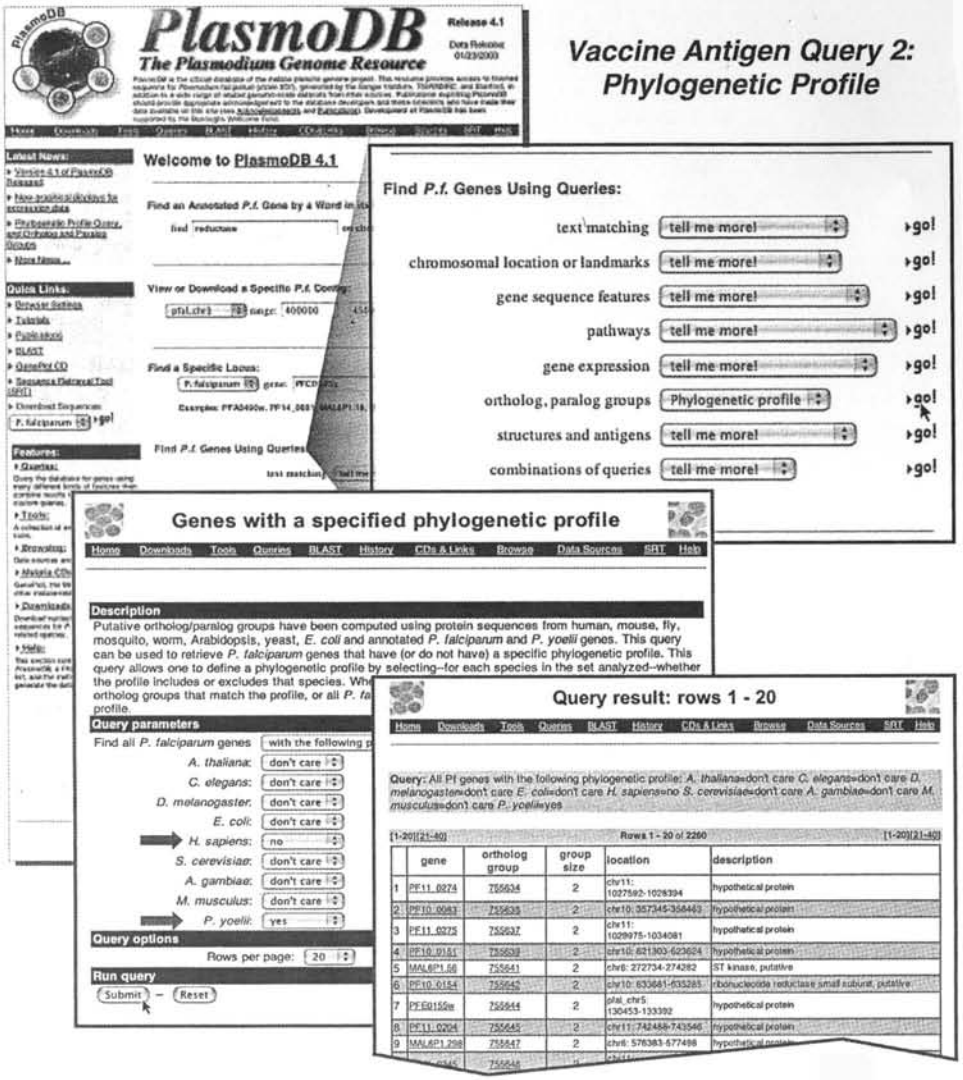

Figure 4. Phylogenomic cross-comparisons with other genome sequence data can identify putative orthologous genes (Li *et al.*, 2003b), and genes that are phylogenetically-restricted in their distribution (Ajioka *et al.*, 1998). In seeking candidate vaccine targets, one might wish to identify antigens that are highly conserved between *P. falciparum* and *P. yoelii*, but not shared with the human host. This query yields 2260 hits (>40% of the parasite genome).

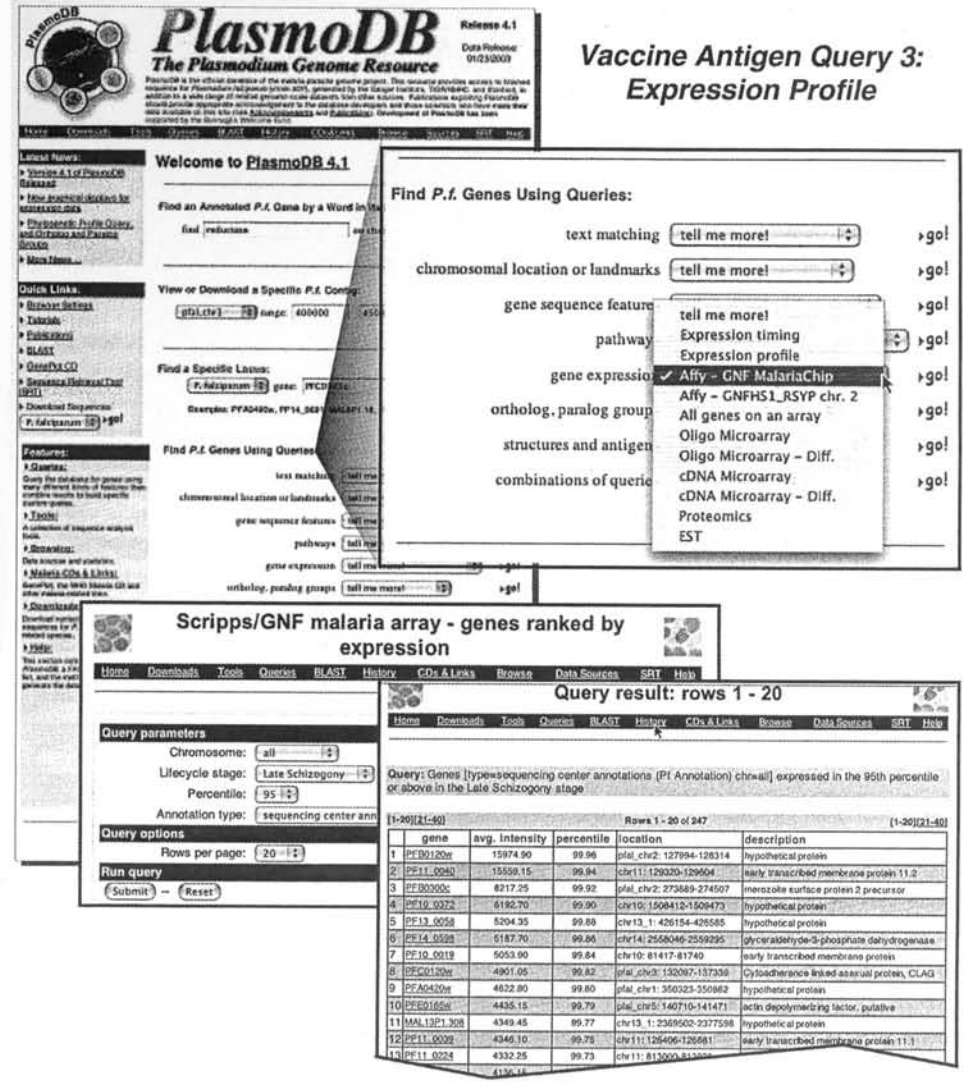

Figure 5. In seeking a blood-stage vaccine, one might wish to prioritize antigens that are abundantly expressed in the extra-erythrocytic merozoite stage. This query focuses on experiments conducted by Le Roch *et al.* (2003b) using an Affymetrix microarray to examine expression across the erythrocytic life cycle, and seeks genes that are among the top 5% in steady-state transcript abundance during late schizogony (many other expression datasets and query strategies can also be envisioned).

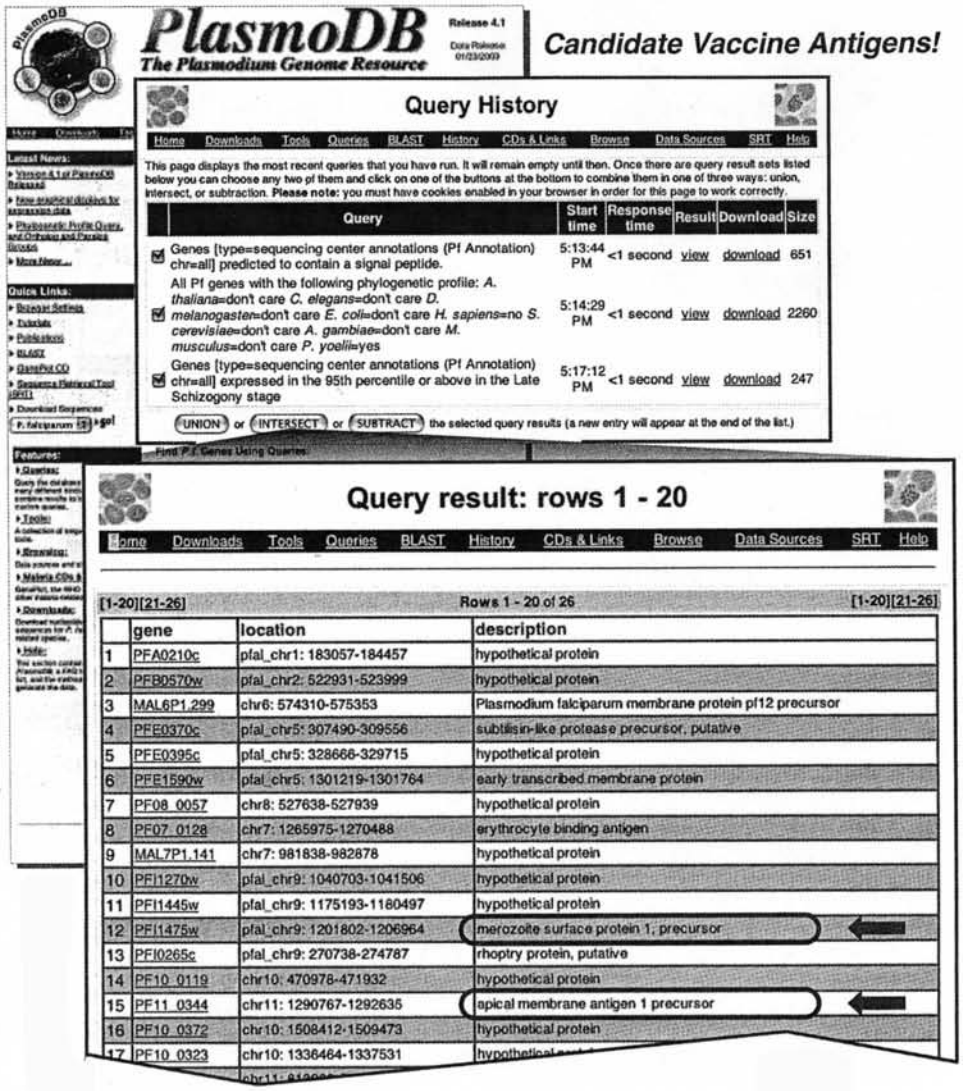

Figure 6. Using the 'Query History' feature of PlasmoDB to combine the queries illustrated in Figures 1-3 identifies only 26 genes exhibiting all three desired characteristics: antigens that are secreted, restricted to *Plasmodium* species, and abundantly transcribed just before merozoite emergence. Among these genes are both of the leading erythrocytic vaccine candidates: MSP1 and AMA1.

P. yoelii, but absent from the human genome, as shown in Figure 4. One might further wish to restrict consideration to abundant transcripts expressed in late schizonts, based on the GNF photolithographic array, as shown in Figure 5. Each of these queries yields a list of several hundred (or thousand) genes, but exploiting the "History" function of PlasmoDB permits taking the intersection of these queries, yielding 26 hits (Figure 6) … including two of the leading vaccine antigens now undergoing trials (MSP1 and AMA1). The remaining proteins (mostly annotated as hypothetical proteins) would be interesting to explore as candidate vaccine antigens.

Of course, there are many other ways to configure this search, including refining the desired expression pattern, considering chromosomal location, evaluating potential efficacy, or seeking for evidence for positive selection from population genetic studies, etc (although the data is not yet in place for all of these queries). The point of this exercise is not that analysis *in silico* is ever likely to take the place of laboratory analysis (particularly in the case of vaccine antigen discovery!) Rather, the point is that computational tools can rapidly filter available options, providing each gene in the dataset with a set of credentials that can be assessed for potential vaccine efficacy. Overall, the goal is to let computers do what computers do well (integrating and analyzing large-scale datasets), and let people do what people do well (experimental validation at the laboratory bench).

4.1. Future Directions

Plasmodium bioinformatics resources are growing daily, and we can anticipate the incorporation of new data at an ever-accelerating rate. The year following completion of reference sequences for *P. falciparum* and *A. gambiae* saw the release of effectively complete genome sequence for *P. yoelii* (and other apicomplexan parasite species), extensive sequence information for several other *Plasmodium* species, whole-genome RNA and protein expression data (on several platforms; Florens *et al.*, 2003; Le Roch *et al.*, 2002; 2003; Bozdech *et al.*, 2003a; 2003b), and new algorithms for cross-genome comparisons (Li *et al.*, 2003b).

The coming year is likely to bring further sequence data for a field isolate of *P. falciparum* and additional *Plasmodium* species (and other apicomplexan parasites); revised and updated annotation for *P. falciparum* and *P. yoelii;* next-generation computational analyses of these genomes, incorporating new algorithms for orthologous group identification; syntenic analysis and other comparisons across species boundaries; genome-wide SNP markers for genetic studies; incorporation of greatly explanded EST datasets, representing several life-cycle stages; additional transcript profiling data, including studies on additional strains, life cycle stages, and treatments; additional proteomics data, including quantitative data from various life cycle stages, and preliminary analysis of protein modifications.

In future, we can also anticipate the availability of additional data types, from population genetic data, to clinical records, to structural genomics results, to publications records. The computational challenge will be to integrate these emerging data types with existing database resources, and develop analysis tools for effective database mining. The biological challenge is to consider how to effectively translate biological questions into computationally accessible terms. What questions do ***you*** want to ask?

- What features characterize potential drug targets, vaccine antigens, diagnostics?
- How to get a handle on understudied life cycle stages?
- How to map virulence genes and other loci of interest?
- What features define parasite proteins likely to interact with the red cell, liver, and host endothelium?
- What genes to target for genetic knock-outs, knock-downs, etc?
- How best to explore parasite population biology?
- How best to compare *P. falciparum* with other *Plasmodium*, apicomplexan, and other eukaryotic pathogen species: gene families, taxonomically- or functionally-restricted genes?
- What information can we usefully extract from the *P. vivax* genome?
- How to exploit genomic information for host and vector species (*Plasmodium* vs. human, mouse, *Anopheles*)?
- What information is of greatest interest for studying eukaryotic biology and evolution?
- How best to link genomics data to publication records?
- How to integrate clinical data?
- What new 'omics'-scale datasets would be useful?

5. Concluding Comments: Bioinformatics Research, and Where to Look for Further Assistance

Now is an exciting time to be engaged in malaria research. The availability of genome sequences for the parasite host and vector, along with emerging expression analyses and anticipated population data, are providing unprecedented insight into the biology of *Plasmodium* and its interaction with its hosts. Utilization of this data requires proper storage, retrieval, analysis and integration of these and new data types.

Databases offer a tremendous asset for biomedical research, but they do not obviate the need for critical thinking; the same analytical approach is required for bioinformatics experiments conducted *in silico* as for experimental work conducted at the laboratory bench. Problems are likely to arise whenever the exact nature of the data type or bioinformatics analysis tool is not understood. Database users should therefore endeavor to fully explore database resources, and should not be reluctant to contact the database developers whenever questions or problems arise that are not clearly explained in the documentation provided.

Bioinformatics assistance is readily available in multiple forms. Many of the major databases provide tutorial, "How To" and "Frequently Asked Questions" help pages. A few minutes spent reading this material can save hours of frustration or misuse/misinterpretation of the data contained in the database. If questions or doubts still remain, contact the database directly via e-mail. Clearly state your question(s), referring to the exact pages or tools, your computer platform (Mac, Windows, Unix) and browser type and version. Many "bugs" are platform- or browser-specific. Explanations of specialty databases can be found in the annual Nucleic Acids Research database issue published each January. Tutorials and in-depth explanations of analysis tools can usually be found on the tool's web site or in bioinformatics books. Classes pertaining to the use of malaria related resources are routinely offered by the Malaria Research and Reference Reagent Resource Center (MR4 - http://www.malaria.mr4.org) and the WHO/TDR (http://www.who.int/tdr/). Workshops on how to use malaria-related databases are offered at several international meetings (Molecular Parasitology, Tropical Medicine and others), participants are encouraged to check meeting agendas and contact meeting organizers.

Finally, it is important to keep in mind that computational analysis is accessible to everyone ... you can do it! The advantage of bioinformatics research relative to most bench work is that computational experiments can often be run quickly, at negligible cost, and with no risk of damaging the starting material or wasting reagents. It should also be noted that there are usually many, many routes to an answer: many approaches to predicting genes, many methods for defining protein features, many expression datasets, many methods for analyzing expression data, etc. As long as the raw data is available, new analyses and re-analysis can be and should be performed as new techniques and experimental strategies develop.

6. Acknowledgements

The authors would like to thank the numerous researchers and students who have generated genomic-scale datasets and made these resources publicly available. We would particularly like to thank the many malaria researchers whose questions – whether submitted electronically, during workshops and training sessions, or in the laboratory – have contributed to the success of PlasmoDB and other databases. We also thank members of our laboratories, Bindu Gajria, Philip Labo, and Boris Striepen for useful comments on this manuscript.

7. References

Adams, J. H., Wu, Y., and Fairfield, A. 2000. Malaria Research and Reference Reagent Resource Center. Parasitol. Today. 16: 89.

Ajioka, J. W., Boothroyd, J. C., Brunk, B. P., Hehl, A., Hillier, L., Manger, I. D., Marra, M., Overton, G. C., Roos, D. S., Wan, K. L., Waterston, R., and Sibley, L. D. 1998. Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. Genome Res. 8: 18-28.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 25: 25-29.

Bahl, A., Brunk, B., Coppel, R. L., Crabtree, J., Diskin, S. J., Fraunholz, M. J., Grant, G. R., Gupta, D., Huestis, R. L., Kissinger, J. C., Labo, P., Li, L., McWeeney, S. K., Milgram, A. J., Roos, D. S., Schug, J., and Stoeckert, C. J., Jr. 2002. PlasmoDB: the *Plasmodium* genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). Nucl. Acids Res. 30: 87-90.

Baxevanis, A. D. 2003. The Molecular Biology Database Collection: 2003 update. Nucl. Acids Res. 31: 1-12.

Baxevanis, A. D., Davison, D. B., Page, R. D. M., Petsko, G., Stein, L., and Stormo, G. D. 2003. Current Protocols in Bioinformatics. Rockville.

Baxevanis, A. D., and Ouellette, B. F. 2001. Bioinformatics: A Paractical Guide to the Analysis of Genes & Proteins. Wiley-Interscience, New York.

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. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature. 400: 532-538.

Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., and DeRisi, J. L. 2003a. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol. 4: R9.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. 2003b. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 1: 85: 100.

Carlton, J., and Dame, J. 2000. The *Plasmodium vivax* and *P. berghei* gene sequence tag projects. Parasitol. Today 16: 409.

Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., van Aken, S. E., Riedmuller, S. B., Feldblyum, T. V., Cho, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L. M., Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J., and Carucci, D. J. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature. 419: 512-519.

Davidson, S. B., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, G. C., and Stoeckert, C.J. 2001. K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. IBM Syst. J. 40: 512-531

del Portillo, H. A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C. P., Schneider, N. K., Villalobos, J. M., Rajandream, M. A., Harris, D., Pereira da Silva, L. H., Barrell, B., and Lanzer, M. 2001. A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. Nature. 410: 839-842.

Ferdig, M. T., and Su, X. Z. 2000. Microsatellite markers and genetic mapping in *Plasmodium falciparum*. Parasitol. Today 16: 307-312.

Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., and Carucci, D. J. 2002. A proteomic view of the *Plasmodium falciparum* life cycle. Nature. 419: 520-526.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. 2002a. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 419: 498-511.

Gardner, M. J., Shallom, S. J., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L., and Fraser, C. M. 2002b. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. Nature. 419: 531-4.

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G. G., Clayton, R., White, O., Smith, H. O., Fraser, C. M., Hoffman, S. L., *et al.* 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science. 282: 1126-1132.

Gibson, G., and Muse, S. 2001. A Primer of Genome Science. Sinauer, Sunderland.

Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S., Barron, A., Brooks, K., Buckee, C. O., Burrows, C., Cherevach, I., Chillingworth, C., Chillingworth, T., Christodoulou, Z., Clark, L., Clark, R., Corton, C., Cronin, A., Davies, R., Davis, P., Dear, P., Dearden, F., Doggett, J., Feltwell, T., Goble, A., Goodhead, I., Gwilliam, R., Hamlin, N., Hance, Z., Harper, D., Hauser, H., Hornsby, T., Holroyd, S., Horrocks, P., Humphray, S., Jagels, K., James, K. D., Johnson, D., Kerhornou, A., Knights, A., Konfortov, B., Kyes, S., Larke, N., Lawson, D., Lennard, N., Line, A., Maddison, M., McLean, J., Mooney, P., Moule, S., Murphy, L., Oliver, K., Ormond, D., Price, C., Quail, M. A., Rabbinowitsch, E., Rajandream, M. A., Rutter, S., Rutherford, K. M., Sanders, M., Simmonds, M., Seeger, K., Sharp, S., Smith, R., Squares, R., Squares, S., Stevens, K., Taylor, K., Tivey, A., Unwin, L., Whitehead, S., Woodward, J., Sulston, J. E., Craig, A., Newbold, C., and Barrell, B. G. 2002. Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. Nature. 419: 527-531.

Hayward, R. E., Derisi, J. L., Alfadhli, S., Kaslow, D. C., Brown, P. O., and Rathod, P. K. 2000. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. Mol. Microbiol 35: 6-14.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., Atkinson, P. W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chatuverdi, K., Christophides, G. K., Chrystal, M. A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C. A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M. E., Hladun, S. L., Hogan, J. R., Hong, Y. S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J. J., Lobo, N. F., Lopez, J. R., Malek, J. A., McIntosh, T. C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S. D., O'Brochta, D. A., Pfannkoch, C., Qi, R., Regier, M. A., Remington, K., Shao, H., Sharakhova, M. V., Sitter, C. D., Shetty, J., Smith, T. J., Strong, R., Sun, J., Thomasova, D., Ton, L. Q., Topalis, P., Tu, Z., Unger, M. F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K. J., Wortman, J. R., Wu, M., Yao, A., Zdobnov, E. M., Zhang, H., Zhao, Q., *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science. 298: 129-149.

Hyman, R. W., Fung, E., Conway, A., Kurdi, O., Mao, J., Miranda, M., Nakao, B., Rowley, D., Tamaki, T., Wang, F., and Davis, R. W. 2002. Sequence of *Plasmodium falciparum* chromosome 12. Nature. 419: 534-537.

Janssen, C. S., Barrett, M. P., Lawson, D., Quail, M. A., Harris, D., Bowman, S., Phillips, R. S., and Turner, C. M. R. 2001. Gene discovery in *Plasmodium chabaudi* by genome survey sequencing. Mol. Biochem. Parasitol. 113: 251-260.

Kissinger, J. C., Brunk, B. P., Crabtree, J., Fraunholz, M. J., Gajria, B., Milgram, A. J., Pearson, D. S., Schug, J., Bahl, A., Diskin, S. J., Ginsburg, H., Grant, G. R., Gupta, D., Labo, P., Li, L., Mailman, M. D., McWeeney, S. K., Whetzel, P., Stoeckert, C. J., and Roos, D. S. 2002. The *Plasmodium* genome database. Nature 419: 490-492.

Lai, Z., Jing, J., Aston, C., Clarke, V., Apodaca, J., Dimalanta, E. T., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T. S., Paxia, S., Hoffman, S. L., Craig Venter, J., Huff, E. J., and Schwartz, D. C. 1999. A shotgun optical map of the entire *Plasmodium falciparum* genome. Nature Genet. 23: 309-313.

Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M., Pain, A., Sauerwein, R. W., Eling, W. M., Hall, N., Waters, A. P., Stunnenberg, H. G., and Mann, M. 2002. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature. 419: 537-542.

Le Roch, K. G., Zhou, Y., Batalov, S., and Winzeler, E. A. 2002. Monitoring the chromosome 2 intraerythrocytic transcriptome of *Plasmodium falciparum* using oligonucleotide arrays. Am. J. Trop. Med. Hyg. 67: 233-243.

Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., and Winzeler, E.A. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. 301: 1503-1508.

Li, L., Brunk, B. P., Kissinger, J. C., Pape, D., Martin, J., Wylie, T., Dante, M., Tang, K., Cole, R., Fogarty, S. J., Howe, D. K., Liberator, P. A., Diaz, C., White, M.,

Jerome, M. E., Johnson, E. A., Radke, J. A., Waterston, R., Clifton, S., Roos, D. S., and Sibley, L. D. 2003a. Gene Discovery in the Apicomplexa as Revealed by EST Sequencing and Assembly of a Comparative Gene Database. Genome Res. 13: 443-454.

Li, L., Stoeckert, C. J., and Roos, D. S. 2003b. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178-2190.

Mamoun, C. B., Gluzman, I. Y., Hott, C., MacMillan, S. K., Amarakone, A. S., Anderson, D. L., Carlton, J. M.-R., Dame, J. B., Chakrabarti, D., Martin, R. K., Brownstein, B. H., and Goldberg, D. E. 2001. Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. Mol. Microbiol. 39: 26-36.

Mount, D. 2001. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Mu, J., Duan, J., Makova, K. D., Joy, D. A., Huynh, C. Q., Branch, O. H., Li, W. H., and Su, X. Z. 2002. Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. Nature. 418: 323-326.

Munasinghe, A., Patankar, S., Cook, B. P., Madden, S. L., Martin, R. K., Kyle, D. E., Shoaibi, A., Cummings, L. M., and Wirth, D. F. 2001. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. Mol. Biochem. Parasitol. 113: 23-34.

Pleasance, E. D., Marra, M., and Jones, S. J. M. 2003. Assessment of SAGE in Transcript Identification. Genome Res. 13: 1203-1215.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. P. 2001. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucl. Acids Res. 29: 159-164.

Schomburg, I., Chang, A., and Schomburg, D. 2002. BRENDA, enzyme data and metabolic information. Nucl. Acids Res. 30: 47-49.

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B. P., and Stoeckert, C. J., Jr. 2002. Predicting gene ontology functions from ProDom and CDD protein domains. Genome Res. 12: 648-655.

Su, X., Ferdig, M. T., Huang, Y., Huynh, C. Q., Liu, A., You, J., Wootton, J. C., and Wellems, T. E. 1999. A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. Science. 286: 1351-1353.

Tchavtchitch, M., Fischer, K., Huestis, R. L., and Saul, A. 2001. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reaveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. Mol. Biochem. Parasitol. 118: 211-222.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. 1995. Serial analysis of gene expression. Science 270: 484-487.

Watanabe, J., Sasaki, M., Suzuki, Y., and Sugano, S. 2001. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*. Nucl. Acids Res. 29: 70-71.

Wootton, J. C., Feng, X., Ferdig, M. T., Cooper, R. A., Mu, J., Baruch, D. I., Magill, A. J., and Su, X. Z. 2002. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature. 418: 320-323.