THE UNIVERSITY *of* EDINBURGH

Visualising Plasmodium falciparum functional genomic data in MaGnET: Malaria Genome Exploration Tool

Joanna Louise Sharman

A thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2009

Unless otherwise stated (see "Declarations and Acknowledgements" following each chapter), the work described in this thesis is my own work and has not been submitted in whole or in part for a degree or other qualification at this, or any other university.

Joanna L. Sharman

University of Edinburgh

2009

ABSTRACT

Malaria affects the lives of 500 million people around the world each year. The disease is caused by protozoan parasites of the genus *Plasmodium*, whose ability to evade the immune system and quickly evolve resistance to drugs poses a major challenge for disease control. The results of several *Plasmodium* genome sequencing projects have revealed how little is known about the function of their genes (over half of the approximately 5400 genes in *Plasmodium falciparum,* the most deadly human parasite, are annotated as 'hypothetical'). Recently, several large-scale studies have attempted to shed light on the processes in which genes are involved; for example, the use of DNA microarrays to profile the parasite's gene expression.

With the emergence of varied types of functional genomic data comes a need for effective tools that allow biologists (and bioinformaticians) to explore these data. The goal of exploration/browsing-style analyses will typically be to derive clues towards the function of thus far uncharacterised gene products, and to formulate experimentally testable hypotheses. Graphic interfaces to individual data sets are obviously beneficial in this endeavour. However, effective visual data exploration requires also that interfaces to different functional genomic data are integrated and that the user can carry forward a selected group of genes (not merely one at a time) across a variety of data sets. Non-expert users especially benefit from workbench-like tools offering access to the data in this way. Still, only very few of the contemporary publicly available software have implemented such functionality.

This work introduces a novel software tool for the integrated visualisation of functional genomic data relating to *P. falciparum*: the Malaria Genome Exploration Tool (MaGnET).

MaGnET consists of a light-weight Java program for effective visualisation linked to a MySQL database for data storage. In order to maximise accessibility, the program is publicly available over the World Wide Web (http://www.malariagenomeexplorer.org/). MaGnET incorporates a Genome Viewer for visualising the location of genomic features, a Protein-Protein Interaction Viewer for visualising networks of experimentally determined interactions and an Expression Data Viewer for displaying mRNA and protein expression data. Complex database queries can easily be constructed in the Data Analysis Viewer. An advantage over most other tools is that all sections are fully integrated, allowing users to carry selected groups of genes across different datasets. Furthermore, MaGnET provides useful advanced visualisation features, including mapping of expression data onto genomic location or protein-protein interaction network. The inclusion of available third-party Java software has expanded the visualisation capability of MaGnET; for example, the Jmol viewer has been incorporated for viewing 3-D protein structures.

An effort has been made to only include data in MaGnET that is at least of reasonable quality. The MaGnET database collates experimental data from various public *Plasmodium* resources (e.g. PlasmoDB) and from published functional genomic studies, such as DNA microarrays. In addition, through careful filtering and labelling we have been able to include some predicted annotation that has not been experimentally confirmed, such as Gene Ontology and InterPro functional assignments and modelled protein structures.

The application of MaGnET to malaria biology is demonstrated through a series of small studies. Initial examples show how MaGnET can be used to effectively demonstrate results from previously published analyses. This is followed

up by using MaGnET to make a set of predictions about the possible functions of selected uncharacterised genes and suggesting follow-up experiments.

ACKNOWLEDGEMENTS

Funding: Thanks to the Medical Research Council for funding this PhD project. Thanks to the University of Edinburgh James Rennie Bequest Fund for funding me to attend the BioMalPar Conference in Heidelberg, 2005. Thanks to Dietlind Gerloff for sponsoring a visit to the University of California, Santa Cruz, 2007.

PhD supervision: Many thanks to Dietlind for her continued encouragement, support, guidance and enthusiasm throughout this project. Special thanks for her continued involvement after she moved to California. Thanks to Simon Tomlinson for "taking over" from Dietlind as my Edinburgh-based supervisor, and his constructive advice and encouragement. Thanks to my thesis committee for their helpful comments.

Colleagues and technical support: Thanks to everyone in Swann 3 and Darwin 2 for making it a pleasant and fun place to work. Thanks to Thomas Juettemann for computing support and to Paul Taylor for infrastructure support. Thanks to UCSC School of Engineering for hosting the software and database.

Family and friends: Last, but certainly not least, thanks to my family and friends for always believing in me. I would like to especially thank my parents, Jack and Gill, for their love, support, patience, advice and encouragement. Many thanks to my grandparents, Tim and Margaret, whose encouragement while I was writing this thesis was much appreciated! Thanks to my sister, Harri, my "in-laws" Vilma, Divya, Jitesh and Rohan, and to all my friends, especially Jen, Thea, Laura, Marie, Karen and Becky, for all the laughs, love and support. Finally, a million thanks to my husband, Dinesh, for everything, and for being an inspiration!

Also, thanks to the many people who have provided useful comments that influenced the design of the software described in this thesis. All scientific acknowledgements follow each chapter.

CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF TABLES

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Chapter 6

ABBREVIATIONS

A + T	Adenine and Thymine
API	Application Programming Interface
ATP	Adenosine triphosphate
AWT	Abstract Window Toolkit
bp	base pairs
BLAST	Basic Local Alignment Search Tool
CAK	CDK-activating kinase
CATH	Class (C), Architecture (A), Topology (T) and Homologous superfamily (H)
CDK	cyclin-dependent kinases
CDT	Clustered Data Table
CHD	chromodomain-helicase-DNA-binding protein
CRK	CDK-related kinases
CTD	carboxyl-terminal domain
CLAG	cytoadherence linked asexual protein
E	Expectation value
EBA	erythrocyte binding antigen
EMBL	European Molecular Biology Laboratory
EMP	erythrocyte membrane protein
EST	Expressed Sequence Tag
G + C	Guanine and Cytosine
GO	Gene Ontology
GPL	Gnu General Public Licence
GUI	Graphical User Interface
HSP	heat shock protein
HTML	HyperText Markup Language
IDC	Intraerythrocytic Developmental Cycle
IE	Infected Erythrocyte
IEA	Inferred from Electronic Annotation
ISS	Inferred from Sequence or Structural Similarity
JDBC	Java Database Connectivity
JDK	Java Development Kit
JRE	Java Runtime Environment
JVM	Java Virtual Machine
KAHRP	knob associated histidine rich protein
kb	kilobases
KEGG	Kyoto Encyclopaedia of Genes and Genomics
KO	knock-out
LGPL	GNU Lesser General Public License
MaGnET	Malaria Genome Exploration Tool
mb	megabases
MudPIT	multidimensional protein identification technology
NCBI	National Center for Biotechnology Information

OBO	Open Biomedical Ontologies
ORF	open reading frame
OPI	Ontology-based Pattern Idenentification
PDB	Protein Data Bank
PGDB	Pathway/Genome Database
pir	*Plasmodium* interspersed repeats
PVM	parasitophorous vacuolar membrane
RAM	Random Access Memory
RDBMS	relational database management systems
RESA	ring-infected erythrocyte surface antigen
REX	ring exported
RH	reticulocyte binding protein homolog
RIFIN	repetitive interspersed family
SAGE	Serial Analysis of Gene Expression
SCOP	Structural Classification of Proteins
SGPP	Structural Genomics of Pathogenic Protozoa
Skp	S-phase kinase-associated protein
SNP	single nucleotide polymorphism
STEVOR	subtelomeric variable open reading frame family
SQL	Structured Query Language
SURFIN	surface-associated interspersed protein
TIGR	The Institute for Genome Research
TSExplorer	Time-series Explorer
UML	Unified Modelling Language
WHO	World Health Organisation
WT	wild type
WTSI	Wellcome Trust Sanger Institute
Y2H	yeast two-hybrid
YETI	Yeast Exploration Tool Integrator

1. INTRODUCTION

Overview

The post-genomic era has brought new opportunities for studying gene functions of the malaria-causing parasite, *Plasmodium falciparum*. Recent advances in functional genomic research that make use of the parasite's genome sequence have generated a wealth of data about various aspects of gene function and expression. The challenge now is to provide computational tools that assist the malaria research community to mine the data and forge hypotheses that direct experimental research.

In this PhD thesis I present a novel computational tool that meets this need by facilitating researchers to explore functional genomic data about *P. falciparum*. I will describe examples of applying the tool to generate experimentally testable hypotheses of interest to malaria biologists.

1.1 A brief history of malaria and its current status

Malaria is a global disease with ancient roots. The history of mankind's affliction with malaria is long and well documented and this disease is probably responsible for more deaths and suffering than any other (Malaria Site; http://www.malariasite.com).

During the 19th and early 20th Centuries the geographical extent of malaria was so great that over half of the world's population were at risk. Improvements in health and living conditions and the first world-wide control efforts put in place by the World Health Organisation (WHO) in the 1950s saw the eradication of malaria in temperate and seasonal regions and reduced its prevalence in several tropical areas

Centers for Disease Control and Prevention - Malaria; http://www.cdc.gov/malaria/; Malaria Site; http://www.malariasite.com). However, in Africa the eradication program was largely unsuccessful and despite an initial decline in malaria-related deaths, by the late 1980s this trend had reversed (Carter and Mendis 2002). A number of social, economical, environmental and political circumstances contributed to the failure of the strategy in Africa; the greatest cause was probably the emergence and rapid spread of drug resistance among parasites to the widely used anti-malarial, chloroquine. Finally, amid renewed outbreaks in other parts of the world, the eradication hope was abandoned and downgraded to that of control (Carter and Mendis 2002).

Today, the WHO estimates that around 40% of the world's population is at risk from contracting malaria. Each year, over 500 million people become ill with malaria and of these, 1-3 million die. The majority of cases (90%) occur in sub-Saharan Africa and most fatalities in this region are young children and pregnant women. Tropical and sub-tropical regions of the world are also affected, including South and Central America, South East Asia and the Middle East (Malaria Site; http://www.malariasite.com; World Health Organisation Malaria Fact Sheet; http://www.who.int/mediacentre/factsheets/fs094/en/index.html).

In recent years there has been a renewed effort to combat malaria with new strategies and emerging technologies. The driving forces have been partnerships, such as the 'Roll Back Malaria Partnership' (http://www.rollbackmalaria.org/), bringing together governments, health care facilities and charities with academic and commercial scientific research organisations. Focussing on integration of resources within the local health care network and encouraging community responsibility are

essential to maintain control over malaria. These campaigns have centred on the most important tools for malaria control – drugs to treat the disease and methods of reducing human contact with the mosquito vector (Carter and Mendis 2002). For example, spraying of houses and the use of insecticide-impregnated bed nets to reduce transmission of the parasite has been shown to be effective in reducing cases of malaria and child mortality (Greenwood et al. 2005).

In 2002, the Malaria Genome Sequencing Consortium published the first genome sequences of the parasite (Carlton et al. 2002; Gardner et al. 2002) alongside that of its vector, the *Anopheles* mosquito (Holt et al. 2002). With the human genome sequence already published (Lander et al. 2001; Venter et al. 2001), it was hoped that knowledge of these genome sequences would lead to a greater understanding of the parasite and interaction with its host and vector, and ultimately to the discovery of new drug and vaccine targets (Wirth 2002).

Indeed, biomedical research to develop novel drugs both to treat malaria and to be used as anti-malarials has since been moving swiftly. The preferred method of malaria treatment is with one of several combinations of drugs on the market, the exact combination that is effective depends on the geographical area and the type of malaria. However, many of these are too expensive for the African market, which remains in need of cheap, readily-available and sustainable drugs. Good management systems for the delivery of treatments are necessary to prevent over- or under-prescription to ensure that drugs achieve their potential and that resistance does not spread (Greenwood et al. 2005).

Malaria vaccine research has also seen positive results. A number of antigens have been identified as eliciting protective immunity or being critical to the function

of the parasite. Several vaccines targeting various stages of the parasite's life cycle are under clinical trials. However, the problem of antigenic polymorphism means that any one vaccine is likely to have only limited efficacy (Greenwood et al. 2005).

1.2 The malaria parasite

Malaria is an infection of the blood caused by parasites of the genus *Plasmodium*. There are over 100 species of *Plasmodia*, infecting many different animals including birds, mammals and reptiles (Centers for Disease Control and Prevention - Malaria; http://www.cdc.gov/malaria/). Plasmodia have a complex life cycle involving an obligate host organism, in which several asexual stages of replication occur (Figure 1.1a). A mosquito vector is required to transmit the parasite and is where the sexual stages of the life cycle occur (Figure 1.1b).

Four species usually infect humans, of which *Plasmodium falciparum* and *Plasmodium vivax* are the most common, with *Plasmodium malariae* being far less predominant than it once was, and *Plasmodium ovale* having the most limited distribution. *P. falciparum* malaria is the most virulent form and is often associated with severe disease and fatality. The disease is characterised by periodic fevers, referred to as "benign tertian" (*P. vivax* and *P. ovale*) – fever every other day, "subtertian, malignant" (*P. falciparum*), or "quartan" (*P. malariae*) – fever every fourth day (Carter and Mendis 2002). The fevers coincide with the release of merozoites from the red blood cell (Figure 1.1). All symptoms of malaria are associated with the blood-stages. *P. falciparum*-infected erythrocytes can adhere to the blood vessel endothelium ('sequestration') [reviewed in (Rogerson 2003)], and may recruit other infected and non-infected red blood cells ('rosetting') [reviewed in

(Rowe 2005)], which is thought to contribute to the severe complication of cerebral malaria. *P. vivax* and *P. ovale* can cause chronic malaria, as parasites can remain latent in the liver for many years (Centers for Disease Control and Prevention - Malaria; http://www.cdc.gov/malaria/; Malaria Site; http://www.malariasite.com).

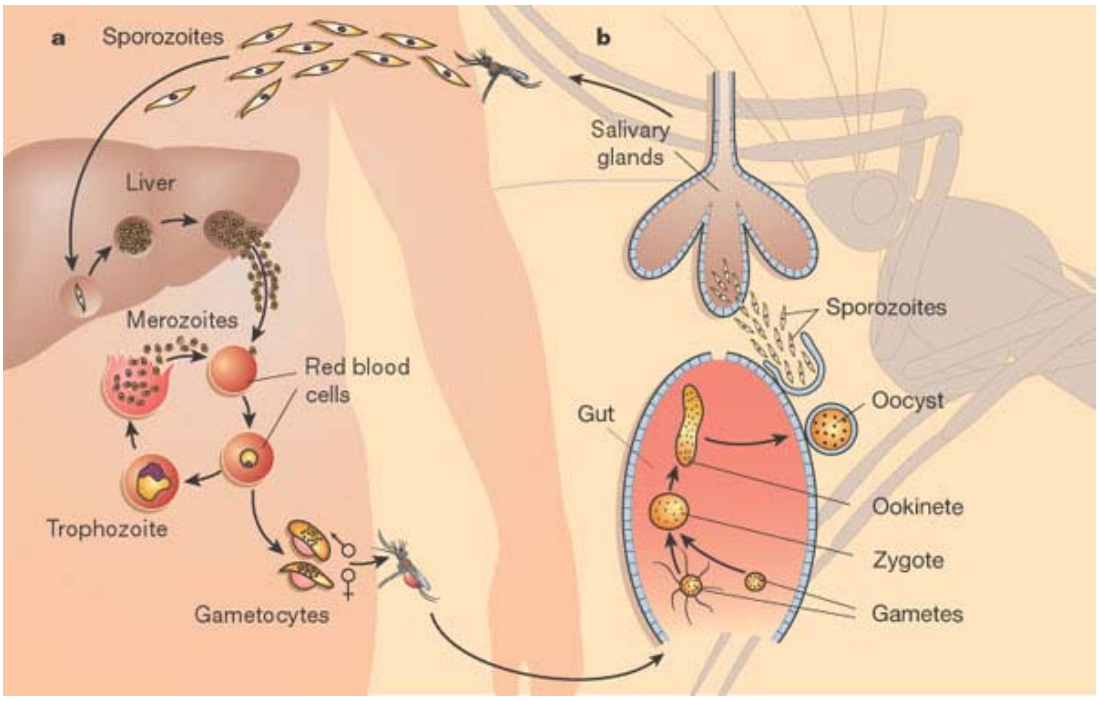


Figure 1.1. Taken from Wirth 2002. Life cycle of the parasite *Plasmodium falciparum*. A, the mosquito injects sporozoites into the blood with its saliva, which quickly invade the liver, where they mature to become merozoites. The merozoites are released into the blood, where they enter erythrocytes and undergo several intra-erythrocytic stages of development, before the host cell bursts to release more merozoites. During this process some parasite cells may become committed to forming male and female gametocytes, the precursors of gametes, which can then be taken up by a feeding mosquito. B, inside the mosquito's stomach the gametocytes mature to gametes, which combine to form a zygote. The zygote develops into an ookinete, which crosses the wall of the gut and becomes an oocyst, filled with sporozoites. When the oocyst bursts, the sporozoites migrate to the salivary glands and the cycle continues.

1.2.1 *Plasmodium* cell biology

Plasmodia are single-celled eukaryotic organisms of the phylum Apicomplexa, which includes related disease-causing organisms *Toxoplasma* and *Cryptosporidium*. Apicomplexa are characterised by a group of organelles known as the 'apical complex' due to their localisation towards one end of the cell. The apical organelles are thought to be involved in the process of host cell invasion. They consist of micronemes, rhoptries and dense granules (heterologous secretory vesicles) (Galinski et al. 2005).

Merozoite invasion of erythrocytes involves four steps:

1. Merozoites bind to the erythrocyte surface by specific interactions occurring between proteins on the parasite's surface and receptors on the erythrocyte.
2. Following binding, the merozoite reorients itself so that the apical end is next to the erythrocyte membrane.
3. The contents of the micronemes are discharged and contribute to formation of a tight junction between the parasite and the host cell. Several microneme proteins that are involved in interactions with receptors on the erythrocyte surface have been identified. In fact, there is some redundancy in these interactions as the parasite can switch between several distinct pathways, which fall into two types: sialic acid-dependent and sialic acid-independent pathways.
4. The parasite actively invades the host cell by first disrupting the erythrocyte cytoskeleton and redistributing its membrane proteins to leave the junction area free. A parasitophorous vacuolar membrane (PVM) forms around the junction, which grows as the parasite enters the erythrocyte. Around this time the rhoptry contents are released and their proteins are associated with development of the

PVM. The PVM is actually believed not to be important for the process of host cell invasion, but is required for the parasite's intracellular development. The junction takes on a ring shape and the parasite moves through it as it enters the parasitophorous vacuole. The force for invasion is generated by unique myosin motors, probably associated with the cytoplasmic region of the parasite ligand proteins, which move along actin filaments and drag the ligand/receptor complexes toward the parasite posterior (Galinski et al. 2005).

The mechanism for other invasive stages of the parasite's life cycle (sporozoites which invade host hepatocytes and ookinetes which invade the mosquito midgut epithelial cells) is slightly different. They often lack rhoptries and do not form a PVM. Sporozoites utilise a gliding motility during invasion, whereby the parasite transmembrane protein thrombospondin-related adhesive protein forms a connection between the substratum and the intracellular myosin motors and is pulled towards the parasite's posterior and deposited on the substratum, forming a trail (Sinden and Gilles 2002).

The final components of the apical complex, dense granules, are released after merozoite entry is complete and are involved in modification of the erythrocyte (Galinski et al. 2005). Several parasite proteins are transported to the host cell membrane and the membrane becomes more permeable to small molecules, needed by the parasite as it grows. Electron-dense protrusions, or 'knobs', appear on the infected erythrocyte (IE) surface. Knobs are thought to be involved in sequestration of the erythrocyte in capillaries. Several proteins are known to be associated with knobs, one of the most important being *P. falciparum* erythrocyte membrane protein 1 (*Pf*EMP1) (Sinden and Gilles 2002). *Pf*EMP1 is encoded by the *var* gene family, a

large polymorphic family implicated in antigenic variation (Baruch et al. 1995; Smith et al. 1995; Su et al. 1995). The specific member that is expressed changes between generations and appears to affect the cytoadherent phenotype (Peters et al. 2002; Horrocks et al. 2005). *Pf*EMP1 is also thought to contribute to rosetting, a trait of some parasite isolates where the IE binds to multiple uninfected erythrocytes (Rowe 2005).

Some surface proteins, such as *Pf*EMP1, are trafficked to the host cell membrane via a unique parasite organelle called the Maurer's cleft, which the parasite sets up in the erythrocyte cytoplasm [reviewed in (Lanzer et al. 2006)]. Maurer's clefts are large, semi-continuous, polymorphic membrane networks stretching from the PVM to the inner face of the erythrocyte plasma membrane. They appear to have multiple functions, including a possible role in signalling and metabolic pathways (Vincensini et al. 2005). Two mechanisms by which proteins may be transported via Maurer's clefts to the erythrocyte surface have been proposed. One model proposes lateral diffusion of membrane-associated proteins through the network, whilst the other suggests that protein-carrying vesicles bud off from the PVM targeted to Maurer's clefts, from where they move on towards the erythrocyte plasma membrane (Lanzer et al. 2006).

1.2.1.1 The apicoplast

Apicomplexa possess an intracellular plastid organelle, which probably arose from an ancestral secondary endosymbiotic event. Secondary endosymbiosis entails a plastid-containing eukaryote being engulfed by another eukaryote, over time shedding the redundant features of the primary eukaryotic host, such as nucleus and cytoplasm, and transferring its proteins to the secondary host nucleus, to leave the

plastid, often surrounded by additional membranes (Ralph 2005; Waller and McFadden 2005).

The *Plasmodium* apicoplast is maintained by its own circular 35-kilobase (kb) DNA chromosome encoding 64 genes, including some open reading frames (ORFs) of unknown function. The function of the apicoplast is largely unknown, although there is now strong evidence for a type II fatty acid biosynthesis pathway, an isoprenoid biosynthesis pathway, and at least part of a heme biosynthesis pathway, occurring here. No doubt, many hundreds of nuclear-encoded proteins are targeted to the apicoplast. Following identification of the sequence features required for targeting, 10% of the nuclear genes were predicted to be targeted to the apicoplast, although the real number is likely to be higher. Moreover, it has become clear that the apicoplast is essential for parasite survival and numerous anti-malarials have been found to target stages of apicoplast expression (Ralph 2005; Waller and McFadden 2005).

1.2.1.2 The mitochondrion

The *Plasmodium* mitochondrion is unusual amongst these organelles. It has one of the smallest genomes yet sequenced: 6 kb encoding just three proteins and several rRNAs. One important mitochondrial function, the tricarboxylic acid cycle, appears to be absent in asexual development stages. However, drugs that disrupt the mitochondrial electron transport chain kill the parasite, so this appears to be an essential process, and ATP generated by oxidative phosphorylation may become important during the later sexual stages. Clues about its possible role in other biochemical pathways have emerged since the nuclear genome sequence allowed proteins targeted to the mitochondria to be predicted. *Plasmodia* mitochondria are

found in close contact with the apicoplast during asexual stages. This arrangement has allowed the parasite to evolve a unique pathway for heme biosynthesis, where some stages occur in the apicoplast and others in the mitochondrion (Vaidya 2005; van Dooren et al. 2006).

1.2.2 The *P. falciparum* nuclear genome

The *P. falciparum* nuclear genome consists of 14 chromosomes totalling about 22.8-megabases (mb). The genome and average gene length is almost twice that of yeast, with over 15% of genes longer than 4 kb (Gardner et al. 2002). Table 1.1 summarises the distinguishing genomic characteristics of *P. falciparum* strain 3D7.

Characteristics	***P. falciparum* 3D7**
Overall (G + C) content (%)	19.4*
No. of genes	5,540[a]
Mean gene length§ (bp)	2,283*
Percentage coding (%)	52.6*
Genes with introns (%)	53.1[a]
No. of exons	13,315[a]
Mean no. of exons per gene	2.4[a]
(G + C) content of exons (%)	23.7*
(G + C) content of introns (%)	13.5*
(G + C) content of intergenic regions (%)	13.6*
Mean length of intergenic regions (bp)	1,694*

*Data from original genome publication
[a]Data from current genome release
§Excluding introns

Table 1.1. Summary of *P. falciparum* 3D7 nuclear genomic characteristics. When the genome sequence was published in 2002 it was still undergoing final stages of finishing and gap closure. As gaps have been closed and gene models have been re-examined based on new evidence the gene content has changed slightly. At the time of writing, the current version of the genome is 2.1.4, released in July 2007 (Wellcome Trust Sanger Institute *Plasmodium falciparum* Genome Projects; http://www.sanger.ac.uk/Projects/P_falciparum/).

Despite the chromosomes varying considerably in length (and even between isolates), the structure of the subtelomeric regions is remarkably conserved within the genome, indicating that the genetic information undergoes extensive recombination over large regions. It is in the subtelomeric regions that many of the genes involved in antigenic variation and interaction with the host reside. The preliminary analysis of the predicted gene functions revealed a relative paucity of certain functional categories, notably that of 'cell cycle', 'cell organisation and biogenesis' and 'transcription factor', whereas others, such as 'physiological processes' and 'cell adhesion', were over-represented. The difference in gene content between this and other sequenced eukaryotes is probably due to a number of reasons, not least its unique parasitic lifestyle, the great evolutionary distances between them and the high (A + T) content of the genome (Gardner et al. 2002).

1.3 Plasmodia genome sequencing projects

Since the genomes of *P. falciparum* strain 3D7 (Gardner et al. 2002) and *Plasmodium yoelii yoelii* (a rodent parasite) (Carlton et al. 2002) were published, several other species have had their genomes sequenced, covering parasites of an array of different organisms, including rodents, primates and birds. There has been a lot of interest in sequencing genomes from other *Plasmodia* because greater understanding of parasite biology and history will come from comparing similar genomes. Many of the selected species have been long used with their respective host organisms as models for studying the disease and testing vaccines and drug responses. Additionally, several other strains of *P. falciparum* are at various stages of sequencing, the knowledge of which will enable the identification of polymorphic

regions that are under selective pressure and might be implicated in, for example, drug resistance. Table 1.2 lists the recent genome sequencing projects and their current status.

Species/strain	Host	Coverage	Status	Sequencing centre	Published
P. falciparum 3D7	Human	>10X	Finished	WTSI, TIGR, Stanford	(Gardner et al. 2002)
P. falciparum Ghanaian Clinical Isolate	Human	8X	Finishing and analysis	WTSI	No
P. falciparum Dd2	Human	>9X	Finishing and analysis	Broad	No
P. falciparum HB3	Human	>10X	Finishing and analysis	Broad	No
P. falciparum IT	Human	1X	Sequencing	WTSI	No
P. falciparum IGH-CR14, JDP8, RAJ116, 87_239	Human	Not known	Initiated	Broad	No
P. vivax Salvador 1	Human	10X	Finished	TIGR	No
P. y. yoelii 17XNL	Rodent	5X	Finished	TIGR	(Carlton et al. 2002)
P. berghei ANKA	Rodent	8X	Finished	WTSI	(Hall et al. 2005)
P. chabaudi AS	Rodent	8X	Finished	WTSI	(Hall et al. 2005)
P. knowlesi H	Primate	8X	Finishing and analysis	WTSI	No
P. reichenowi Oscar	Primate	3X	Sequencing	WTSI	No
P. reichenowi CDC1	Primate	Not known	Initiated	Broad	No
P. gallinaceum 8a	Avian	3X	Finished	WTSI	No

WTSI = Wellcome Trust Sanger Institute, Hinxton, UK
TIGR = The Institute for Genomic Research – now part of J. Craig Venter Institute, Rockville, MD, USA
Stanford = Stanford Genome Technology Center, Palo Alto, CA, USA
Broad = The Broad Institute of Harvard and MIT, Cambridge, MA, USA

Table 1.2. Status of *Plasmodia* genome sequencing projects as of 12/07/2007. Data in this table were compiled from (Wellcome Trust Sanger Institute Protozoan Genomes; http://www.sanger.ac.uk/Projects/Protozoa/; J. Craig Venter Institute Parasite Projects;

http://www.tigr.org/parasiteProjects.shtml#; The Broad Institute Microbial Sequencing Centre - Plasmodium Falciparum Sequencing Project; http://www.broad.mit.edu/seq/msc/; Coppel et al. 2004).

1.4 Genome annotation

Following assembly of the sequenced genome, the next step is to annotate the genomic features. This is usually done automatically by trained gene prediction programs and their output may be checked manually. The task is made easier when there are annotated genomes of related organisms available. Hence, the annotation of *Plasmodia* genomes has improved in speed and accuracy by incorporating information about gene models from the earlier sequenced genomes. Comparing gene models can help to confirm gene exon boundaries which are often difficult to predict computationally (Berry et al. 2004). Unsurprisingly, the initial annotation of the *P. falciparum* and *P. y. yoelii* genomes was much more difficult as there were no closely related genomes available at the time.

The sequencing consortium is committed to a long-term re-annotation of the *P. falciparum* 3D7 genome using the latest automated prediction tools and databases and incorporating new experimental evidence; the results of which are being manually checked. It is hoped that this genome will be a reference on which to base the annotation of newly sequenced genomes (Berry et al. 2004).

1.4.1 Comparative genomics

There is a surprising amount of conservation between the genomes listed in Table 1.2 (Thompson et al. 2001; Kooij et al. 2005). The central regions of the chromosomes contain several syntenic blocks, which have been re-organised through

recombination events. Most of the species-specific genes are contained in the variable subtelomeric regions, at synteny breakpoints and at intrasyntenic indels (Kooij et al. 2005).

An important part of genome annotation is identifying the genes that are orthologous between species and the genes that have arisen by gene duplication events within a species (paralogs) (Thompson et al. 2001). Interestingly, the major differences in gene content among *Plasmodium* species are between genes that interact with the host immune system. Several antigenic gene families have been identified in *Plasmodia*, including a large super-family of antigens (*Plasmodium* interspersed repeats – *pir*) with representatives of varying copy number in each species (Janssen et al. 2004). Gene families that have arisen by gene duplication have been shown to be evolving more rapidly than non-duplicated genes. Therefore, paralogous gene family expansion appears to be significant in the diversification of *Plasmodia*, whereas the function of orthologs is more conserved (Castillo-Davis et al. 2004). Furthermore, identifying orthologs of human parasite genes allows the molecular mechanisms of parasite biology and drug resistance to be investigated in animal models.

Comparative genomics will flag up functional features not only of genes but also intergenic regions. Conserved intergenic regions may be important for gene regulation (Carlton et al. 2005); on the other hand, the absence of particular regulatory motifs from one species may indicate significant differences in gene expression that could contribute to host specificity and speciation (Hall and Carlton 2005).

On an intra-species level, there appears to be a high frequency of local variation between isolates, such as differences in gene copy number, indels and single nucleotide polymorphisms (SNPs). Polymorphisms frequently occur in genes linked to host interaction, drug resistance or strain-specific phenotypes (Kidgell et al. 2006; Llinas et al. 2006).

Therefore, comparing the similarities and differences between *Plasmodia* genomes can indicate which protein-coding genes or non-coding regions are under selective pressure and might be important for drug resistance, pathogenesis, immune evasion or host specificity, among others.

1.4.2 Annotation tools and programs

The first step in genome annotation is to search for homologs in gene databases, usually using an implementation of the BLAST (Altschul et al. 1990) algorithm. Identifying homologs can provide an indication of the gene's likely function, although much care needs to be taken when inferring function from sequence similarity alone because the genes' functions may have diverged. Many programs have been written to predict various structural and functional features about a protein using its primary amino acid sequence. Often these programs use a database of known examples of the feature to look for matches to the query sequence. The strength of the match will be measured against some criteria decided by the authors of the program; usually only the matches that are deemed significant are reported.

The genome sequencing centres use several such programs to facilitate their annotation effort. In addition, many of the publicly available databases have applied their own specific annotation pipelines to the *P. falciparum* genome. Table 1.3

describes a number of electronic annotation tools that have been used to assign putative functions to *P. falciparum* genes and for which results are available to view online.

Tool	Web address	Description	Publication
Gene Ontology	http://www.geneontology.org/	A controlled vocabulary to describe gene and gene product attributes in any organism. Three aspects: biological process, molecular function and cellular component. Provides evidence codes to tag the origin of a term's assignment to a gene.	(Ashburner et al. 2000)
PhyloFacts	http://phylogenomics.berkeley.edu/phylofacts/	An online encyclopaedia and database containing pre-calculated structural, functional and phylogenomic analyses of protein families and domains. Each family is represented by a multiple sequence alignment and hidden Markov model in order to classify user-inputted sequences.	(Krishnamurthy et al. 2006)
InterPro	http://www.ebi.ac.uk/interpro/	A database of protein families, domains and functional sites, integrating major structural and functional databases, such as Pfam, PROSITE, SMART and Superfamily.	(Mulder et al. 2007)
Pfam	http://www.sanger.ac.uk/Software/Pfam/	Pfam is a collection of multiple sequence alignments and hidden Markov models of protein domains and families.	(Finn et al. 2006)
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Predicts the presence and location of signal peptides and their cleavage sites in proteins. Uses neural networks and hidden Markov models.	(Bendtsen et al. 2004)
OrthoMCL	http://orthomcl.cbil.upenn.edu/	A genome-scale algorithm for identifying and groups of gene sequences that are shared across two or more species (orthologs) and groups that represent species-specific gene families (paralogs). Includes a database containing pre-computed results for sequenced genomes.	(Li et al. 2003)
ModBase	http://modbase.compbio.ucsf.edu/	Database of theoretical 3-D protein structure models calculated by comparative modelling. The modelling pipeline uses the PSI-BLAST and MODELLER programs.	(Pieper et al. 2006)

Comparative models of *Plasmodium falciparum* (3D7) proteins	http://bioinfo.icgeb.res.in/codes/model.html	Database of theoretical 3-D protein structure models calculated by comparative modelling. The method used BLASTP and MODELLER.	(Gowthaman et al. 2005)
PlasmoCyc	http://plasmocyc.stanford.edu/ A version also at http://apicyc.apidb.org/	A genome/pathway database for *P. falciparum* 3D7. Graphical display of pathways and individual enzymatic reactions. Provides a whole-cell overview of metabolic pathways and allows comparison to other organisms. Part of the Pathway Tools/BioCyc framework. No longer appears to be supported.	(Yeh et al. 2004)
Kyoto Encyclopaedia of Genes and Genomics (KEGG)	http://www.genome.ad.jp/dbget-bin/www_bfind?p.falciparum	Database of biological systems and their components, including pathways, enzymes, reactions, gene catalogues, gene orthology, and ligands.	(Kanehisa et al. 2006)
Ontology-based Pattern Identification (OPI) Database	http://chemlims.com:8080/OPI20/MServlet.ChemInfo	Portal to search the Winzeler lab's database of the OPI algorithm applied to *P. falciparum* mRNA expression data. OPI identifies expression patterns that best represent existing knowledge of gene function and uses the principal of guilt by association to systematically annotate genes with GO terms.	(Zhou et al. 2005)
PlasmoMAP	http://www.cbil.upenn.edu/plasmoMAP/	Network of predicted functional interactions between *P. falciparum* proteins; constructed by integrating computational and functional genomics data within a Bayesian framework.	(Date and Stoeckert 2006)

Table 1.3. A list of sources of annotation available for *Plasmodium* genes. These tools predict functional annotation using automated electronic methods, usually based on homology to other annotated sequences.

Computational analysis of *P. falciparum* sequences has additionally lead to the predicted reconstruction of metabolic pathways (Yeh et al. 2004; Kanehisa et al. 2006), functional interaction networks (Date and Stoeckert 2006), prediction of protein structures by homology modelling (Gowthaman et al. 2005; Pieper et al. 2006), and a novel theory about the likely mechanism of gene expression control within the parasite (Coulson et al. 2004; van Noort and Huynen 2006; Gunasekera et

al. 2007). The advantages of large-scale homology modelling for protein structure prediction are discussed further in Section 1.5.4.2.

Automated electronic annotation of genes based on sequence homology is a rapid way to achieve an initial indication of their possible biological roles and provide an overview of the genome repertoire. The ultimate aim will be to experimentally confirm all predicted annotations, which, clearly, would take an inordinate amount of time and resources. The next Section will discuss the use of high-throughput functional genomics technologies to help in this endeavour.

1.4.3 Functional classification of gene products

1.4.3.1 Ontologies

Several systems now exist to describe the function of genes using a controlled vocabulary of terms. The most well known and widely used is the Gene Ontology (GO) (Ashburner et al. 2000). The GO project aims to address the need for consistent descriptions of genes across databases and species. The GO project has three hierarchically structured vocabularies (ontologies) to describe genes in terms of their cellular component (C), biological process (P) and molecular function (F). Each GO term has a unique identifying number, a short name and a description; many terms are also associated with synonyms.

A large number of tools have been developed for browsing and annotating genes with GO terms, as well as for mapping GO terms to other types of data, such as clusters of genes derived from microarray analysis of gene expression (for a list of tools see http://www.geneontology.org/). When a term is assigned to a gene, the annotation is required to be given one of a number of GO evidence codes. Evidence

codes indicate how the annotation was made and provide useful information, such as a PubMed article identifier or a database identifier for annotations made using information from other genes, particularly in cases where function is inferred by sequence similarity.

The downside of ontologies is that they aim to be generically applicable to all organisms, so the scope and potential of GO for describing unusual and complex organisms such as *Plasmodium* is somewhat limited. Much of the initial annotation of a genome's gene sequences is inferred automatically using sequence databases to assign putative functions to the genes based on sequence similarity alone. There are a number of problems associated with this method, not least that these programs often rely on local sequence similarity to assign functional annotation, which may be unreliable because the genes' functions may have diverged. As a consequence, many of the genes in the *P. falciparum* genome that have been assigned GO terms, have the evidence code IEA (Inferred from Electronic Annotation), where no curator has checked the annotation, and a smaller number have the evidence code ISS (Inferred from Sequence or Structural Similarity), where a curator has checked the annotation. Furthermore, due to the large number of genes that are unique to the *Plasmodium* genus, annotation of the genome with GO terms was limited to those with a recognisable homolog in the sequence databases [as of 05/07/2008 58% of *P. falciparum* genes have a GO term assigned to them (Wellcome Trust Sanger Institute *Plasmodium falciparum* Genome Projects; http://www.sanger.ac.uk/Projects/P_falciparum/; Malaria Genome Exploration Tool; http://www.malariagenomeexplorer.org)].

1.4.3.2 Databases of protein signatures

Many databases of conserved protein families, domains and functional sites exist to facilitate the annotation of novel protein sequences. Such information is important because patterns of conserved functional sites and domains can provide an indication of a protein's probable function. The InterPro resource integrates ten major protein signature databases (Mulder et al. 2007). The member databases use a variety of methods to annotate protein sequences. Therefore, by combining multiple methods InterPro increases overall sequence annotation coverage compared to that provided by any single method. InterPro has been employed by genome annotation projects as a tool for automated annotation of protein features (Mulder et al. 2007). Indeed, InterPro has been employed by PlasmoDB and GeneDB to annotate genes from *Plasmodium* species [as of 05/07/2008 95% of *P. falciparum* genes have at least one InterPro predicted feature (Wellcome Trust Sanger Institute *Plasmodium falciparum* Genome Projects; http://www.sanger.ac.uk/Projects/P_falciparum/; Malaria Genome Exploration Tool; http://www.malariagenomeexplorer.org)].

1.4.3.3 Novel methods of functional annotation

Recently, attempts have been made to address the short comings of traditional gene sequence annotation methods. One approach that combines the results from a number of different techniques with a conservative attitude to inferring function based on global homology using domain patterns of protein families is PhyloFacts (Krishnamurthy et al. 2006). PhyloFacts is a structural phylogenomic encyclopaedia containing 'books' of protein families and domains, which integrates a variety of experimental and bioinformatic data and methods within an evolutionary context. PhyloFacts was born from the observations that prediction of protein function is

more accurate when using the consensus from a range of different bioinformatic approaches and the power of integrating information from a number of sources. A clear distinction is made between assignments to books representing global homology of sequences (functional annotation can be transferred with high confidence) and those representing single domains or conserved regions (possible function can only be predicted for part of the sequence).

The PhyloFacts project is still undergoing major development and expansion, but it is set to become an important resource for automated genome annotation. This will be enhanced by their forthcoming development of a natural language processing method to create summaries of the key points about each book and the genes assigned to it.

1.5 Functional genomics

The publication of the genome sequence of the malaria parasite with its revelation that two thirds of the genes were marked as 'hypothetical' (predicted genes with no known homologs or with homologs of unknown function), sparked interest in using modern technologies to discover their functions (Doolittle 2002). A number of genome-scale studies involving *Plasmodium* genes and gene products have been published in recent years, including laboratory experiments to measure the levels of gene and protein expression at various stages of the life cycle (Florens et al. 2002; Lasonder et al. 2002; Bozdech et al. 2003; Le Roch et al. 2003; Le Roch et al. 2004; Hall et al. 2005; Llinas et al. 2006), discover protein-protein interactions (LaCount et al. 2005), examine sequence variability of coding and non-coding

regions of the genome (Carret et al. 2005; Kidgell et al. 2006) and characterise protein structures (Mehlin et al. 2006).

In many cases the results of these studies are made available online through publicly accessible databases. The aim of such databases will be to hopefully permit annotations to be updated in future, should new information become available. Table 1.4 summarises the public databases that store experimental data about *Plasmodium* genes and/or proteins.

Tool	Web address	Description	Publication
Protein Data Bank	http://www.rcsb.org/pdb/	Database and information portal of experimentally-solved biological macromolecular structures.	(Berman et al. 2000)
Structural Genomics of Pathogenic Protozoa	http://www.sgpp.org/	Website of the structural genomics initiative for characterisation and structure determination of parasite proteins on a genome-wide scale. Lists the status of target proteins.	(Mehlin 2005; Mehlin et al. 2006)
Malaria Metabolic Pathways Database	http://sites.huji.ac.il/malaria/	Manually curated database of all known pathways in the parasite (metabolic and other processes).	(Ginsburg 2006)
Malaria IDC Transcriptome Database	http://malaria.ucsf.edu/comparison/index.php	Portal to search the DeRisi lab's *P. falciparum* intraerythrocytic developmental cycle (IDC) transcriptome database containing mRNA expression levels as recorded in the microarray experiments described in these papers.	(Bozdech et al. 2003; Llinas et al. 2006)
Malaria Full-Length cDNA database	http://fullmal.ims.u-tokyo.ac.jp/	Online database containing cDNAs from *P. falciparum, P. vivax, P. yoelii and P. berghei*. The aligned sequences can be viewed in a genome browser.	(Watanabe et al. 2004)

Table 1.4. A list of public databases containing experimental data on *Plasmodium* genes and proteins.

In other cases the results of the study have only been made available as a supplementary data file alongside the journal publication. The annotation will remain static and it will be very difficult to track and release updates in future.

The following Sections will provide an overview of the most influential high-throughput functional genomic studies published to date.

1.5.1 Gene expression analysis using DNA microarrays

1.5.1.1 Affymetrix arrays

The Winzeler laboratory (The Scripps Research Institute, La Jolla, CA, USA; http://www.scripps.edu/cb/winzeler/) have used a custom high-density oligonucleotide array produced by Affymetrix (Santa Clara, CA, USA; http://www.affymetrix.com/) to perform whole-genome expression profiling of *P. falciparum* over multiple life cycle stages. The array contains 367,226 probes placed, on average, every 150 bases along both strands of the genome and corresponding to both coding and non-coding regions (including at least 5159 genes). Probes are 25 nucleotides long and genes are each represented by up to 20 probes (the set of probes corresponding to a gene is known as the probe set). The expression level of a gene is estimated by normalising signal intensities across the probe set.

The study by Le Roch et al. (2003) measured the expression profile of *P. falcparum* strain 3D7 over the sporozoite, six intraerythrocytic stages (early and late ring, trophozoite and schizont), merozoite and gametocyte stages of the life cycle. The authors used two different techniques to synchronise the cultures of blood stage parasites and eliminate the effects of chemical or temperature stress from their analysis (5% D-sorbitol treatment and temperature cycling incubation). The results showed that there was little difference in expression between the two methods. The authors found that 4557 genes were expressed in at least one stage of the life cycle,

while 602 genes were not expressed in any stage examined, so these genes might be expressed in the mosquito gut or liver stages (Le Roch et al. 2003).

A later study by Young et al. (2005) measured the expression profile of *P. falciparum* 3D7 gametocyte development on days 1-3, 6, 8 and 12 to capture stages I-V of their development and on days 1-4 for a set of high purity stage I-II gametocytes. They also recorded the expression of *P. falciparum* NF54 gametocytes over 13 days of development. They found that on average 3410 genes were expressed at each time-point of the experiment (Young et al. 2005).

1.5.1.2 Spotted glass slide arrays

The DeRisi laboratory (University of California, San Francisco, CA, USA; http://derisilab.ucsf.edu/) has developed an in-house spotted glass slide microarray containing 7,462 long 70mer oligonucleotides representing 4,488 genes. Of these genes, over 1000 are represented by more than one oligonucleotide. They used a standard two-colour competitive hybridisation mechanism to measure abundances of mRNAs in the cell. During the experiment a pool of cDNAs is assembled from all the timepoints and labelled with Cy3 dye to be used as a reference, while the cDNA from an individual timepoint is labelled with Cy5 dye. For this type of array the analysis does not include normalisation of signal intensities from multiple probes belonging to one gene. This is because there can be differences in the intensities of probes mapping to a single gene, which can arise for a number of reasons, such as splice variation or inaccurate exon prediction.

The DeRisi laboratory used their array to investigate the expression profile of genes in the intraerythrocytic developmental cycle (IDC) of HB3, Dd2 and 3D7 strains of *P. falciparum*. Gene expression was measured hourly for 48, 50 and 53

hours, respectively. The results showed 6415 expression profiles for HB3, 5294 for Dd2 and 6287 for 3D7 (Bozdech et al. 2003; Llinas et al. 2006).

1.5.2 Protein expression analysis using mass spectrometry

Characterisation of the protein content of the parasite's individual life stages is of vital importance to understanding its biology and its interactions with the host and vector. Two studies that were published alongside the parasite's genome sequence characterised the proteome of several stages including sporozoites, asexual stages, gametocytes and gametes. Florens et al. (2002) were able to identify 2,415 proteins over four life stages using multidimensional protein identification technology (MudPIT), which links protein separation by 2D liquid chromatography to tandem mass spectrometry (Florens et al. 2002). Lasonder et al. (2002) used an alternative technique involving protein separation by gel electrophoresis, followed by reverse phase liquid chromatography coupled to quadrupole time-of-flight mass spectrometry, to identify 1,239 proteins in three life cycle stages (Lasonder et al. 2002). More recently, Le Roch et al. (2004) combined the Florens et al. (2002) data with MudPIT analysis of three further life stages in a study that compared the proteome data with that of a previous study of mRNA expression levels (Le Roch et al. 2003) to examine the role of post-transcriptional controls on the regulation of protein expression (Le Roch et al. 2004).

1.5.3 Protein-protein interaction discovery using yeast two-hybrid screening

To date, one high-throughput yeast two-hybrid (Y2H) screen of pair-wise protein-protein interactions has been performed for *Plasmodium* (LaCount et al.

2005). The procedure involves screening a library of 'bait' protein fragments individually against a library of 'prey' protein fragments. The authors were able to identify 2,846 unique interactions between 1,295 different *P. falciparum* proteins. Despite attempting to filter out some of the most likely false positive interactions (by removing protein fragments that interacted with many partners), it is likely that a number of non-biologically meaningful interactions are still represented in the dataset (LaCount et al. 2005; Koegl and Uetz 2007). For instance, two protein molecules that demonstrate the ability to interact physically, may never meet each other under normal physiological circumstances, due to being expressed at different times or in different cellular compartments. Doubtless, other biologically occurring interactions will have been missed within the dataset, because the proteins were not expressed in their native environment. Particularly, secreted, transmembrane and post-translationally modified proteins are unlikely to be picked up by large-scale Y2H experiments (Koegl and Uetz 2007). Other experiments need to be done to produce a 'core' set of repeatable interactions.

1.5.4 High-throughput protein structure characterisation

1.5.4.1 Experimental structure determination using x-ray crystallography

Large-scale studies to clone, express, purify and characterise proteins encoded by the parasite's genome are expected to yield dividends for drug and vaccine target discovery. *Plasmodium* genes are notoriously difficult to express in heterologous expression systems, due to a number of unique features, including the abundance of introns, the AT-rich genome, a different codon bias, proteins

containing long repetitive regions, large disordered loops and an unusual pattern of glycosylation (Mehlin et al. 2006).

The Structural Genomics of Pathogenic Protozoa (SGPP) initiative aims to characterise proteins from major pathogens on a genome-wide scale. Results from an initial selection of 1000 targets resulted in 307 proteins being expressed (although most were insoluble) and just 63 of these with large enough yields to enable further characterisation (Mehlin et al. 2006). So far, several protein structures have been solved from various *Plasmodium* species, including some in complex with ligands (Structural Genomics of Pathogenic Protozoa; http://www.sgpp.org/). In November 2007, the total number of experimental structures in the Protein Data Bank (PDB) (Berman et al. 2000) of *P. falciparum* proteins stood at 137, representing 71 distinct proteins (Malaria Genome Exploration Tool; http://www.malariagenomeexplorer.org).

The structural genomics effort has greatly contributed to understanding the properties that affect a protein's expression in heterologous systems, the knowledge of which will be important for all future functional genomics studies (Mehlin et al. 2006).

1.5.4.2 Structure prediction using homology modelling

Comparative (homology) modelling of tertiary protein structures can be a useful tool for understanding protein function (since structure is directly linked to function) and for development of drugs – for example by using docking programs to screen ligands *in silico*. For an organism such as *Plasmodium*, whose proteins are extremely difficult to express for the reasons given above (Section 1.5.4.1), comparative modelling can increase the number of protein structures available to the

research community. Although models can be greatly improved when taking into account multiple template structures, large-scale modelling efforts are usually based on single template-target alignments, due to the limitations of automated modelling of thousands of proteins (Marti-Renom et al. 2000).

The Database of Comparative Protein Structure Models (ModBase) (Pieper et al. 2006) stores millions of models generated from high-throughput automated modelling of whole-genome sets of predicted proteins, including 48% of *P. falciparum* and 44% of *P. vivax* sequences (Database of Comparative Protein Structure Models; http://modbase.compbio.ucsf.edu/). In a 2005 study Gowthaman *et al*. screened the set of *P. falciparum* 3D7 predicted proteins against the PDB and created models for 476 proteins. They used a cut-off of 40% sequence identity to available template sequence, manually checked the alignment for errors and performed energy minimisation on the model (Gowthaman et al. 2005). The set of models they created are likely to be of high quality, although the overall coverage is relatively low. In contrast, ModBase does not specify any cut-off for sequence identity in order to ensure high coverage and prefers to leave decisions about model quality up to the user [they provide a score for the model using a method that considers various residue-level statistical potentials to assess the fold (Melo et al. 2002)].

1.6 Visualisation and integration of functional genomics data

The results of the annotation effort are made available to the research community in the form of online, publicly available, genome databases. Typically, these databases aim to collate many different types of data in one place and facilitate

users to browse all the characteristics about a gene of interest. To make the process of information retrieval as appealing as possible for biologists, databases often try to incorporate the use of visual tools to display the data.

In the next section purpose-built tools currently available to malaria biologists for browsing and/or analysis of various functional genomic data are discussed. This is followed by an examination of other related tools for functional genomic data organisation and analysis that were not specifically designed for use with *Plasmodium* data. Some of these do provide access to *Plasmodium* datasets, and others are generic tools for use with any organism; however, given the unique properties and problems encountered when dealing with *Plasmodium* data, they are unlikely to work as effective solutions for malaria researchers.

1.6.1 Browsing and analysis tools for *Plasmodium* functional genomic data

Table 1.5 describes several databases and resources dedicated to facilitating users with searching for and displaying information about *Plasmodium* genes.

The main limitation of currently available *Plasmodium* genome databases is that they focus on a single gene-at-a-time view of the data. While being useful resources for biologists wishing to find out information about a particular gene of interest, they do not allow users to effectively explore relationships between genes and to mine functional genomic datasets on a larger-scale. The ability to do this is essential to the discovery of hitherto unknown gene associations with biological processes and for driving the direction of experimental research.

Database	Web address	Description	Publication
PlasmoDB	http://www.plasmodb.org/	Genome database for all sequenced *Plasmodium* species. Includes all kinds of genomic and proteomic data, which is fully searchable. Each gene entry has a page listing information, including homologs, functional annotation, literature, sequence features, recorded expression level, sequences and SNPs, as well as link outs to other databases. Query results can be saved in history and downloaded. Provides a standard genome browser, where pair-wise syntenic regions of certain genomes can be viewed. Aims to be the primary resource for malaria researchers around the world.	(Bahl et al. 2003)
GeneDB	http://www.genedb.org/	Wellcome Trust Sanger Institute (WTSI) Pathogen Genome Database. Holds basic genomic, proteomic and functional data as well as annotation provided computationally and manually about the genes of organisms that have been sequenced at the WTSI. As above, each gene entry has a page listing its information. Users can perform simple searches of the database, and download the results.	(Hertz-Fowler et al. 2004)
	ftp://ftp.sanger.ac.uk/pub/pathogens	The WTSI ftp site where the GeneDB data can be downloaded in flat file format.	
TIGR Parasites Databases	http://www.tigr.org/tdb/parasites/	Gene sequence and annotation database for the genomes sequenced by The Institute of Genomic Research (TIGR): *P. falciparum*, *P. y. yoelii*, and *P. vivax*. Includes GO annotation, metabolic pathways and EST expression data.	N/A
MalariaBase	http://malariabase.org/	A tool for generating potential new annotations for malaria genes, concentrating mainly on *P. falciparum*, *P. y. yoelii* and the mosquito *Anopheles gambiae*. Uses the OntoBlast tool (Zehetner 2003) to assign GO terms from BLAST results. Provides an online database with search facility and ability to store query results and annotate genes. Genes can be analysed for protein family and functional annotation, with link outs to other databases.	N/A
NCBI Malaria Genetics & Genomics	http://www.ncbi.nlm.nih.gov/projects/Malaria/	Resources for *Plasmodium*, including organism-specific BLAST databases, genome and linkage maps, genetic studies, literature, protein structures, SNPs, proteomics data and sequences.	N/A

Broad Institute *P. falciparum* database	http://www.broad.mit.edu/annotation/genome/plasmodium_falciparum_spp/MultiHome.html	Genome sequences, gene predictions and automated annotation generated for *P. falciparum* genomes sequenced by the Broad Institute (and *P. falciparum* 3D7 data from PlasmoDB). Provides genome browsers, BLAST tools and downloadable data files.	N/A
WHO/TDR Malaria Database	http://www.wehi.edu.au/MalDB-www/who.html	Resources for malaria research, including sequence databases, genomic information such as codon usage and restriction sites, information about malaria antigens, ESTs, chromosome maps, literature and links to other sites. Also includes an early genome database (MalDB) for graphical display of chromosomes, gene lists and searching of genomic data (full database downloadable only).	N/A
UCSC *Plasmodium* Genome Browser	http://areslab.ucsc.edu/cgi-bin/hgGateway	A genome browser for viewing conserved regions of several *Plasmodium* genomes, against the predicted *P. falciparum* genes from PlasmoDB. Also shows the locations of ESTs, microarray and mass spec probes used in published mRNA and protein expression studies. Data is available for download and direct access to the MySQL database is possible.	(Chakrabarti et al. 2007)
MalPort	http://malport.bi.up.ac.za:7070/	A portal for bioinformatics databases and tools related to malaria parasites. The SAMP database (Joubert and Joubert 2008) contains computationally-derived predictions for proteins from *P. falciparum*, *P. vivax* and *P. y. yoelii*, such as motifs, secondary and tertiary structure. Also provides lists of candidates for structural characterisation. Advanced query interface provided. MADIBA (Law et al. 2008) provides tools for functional analysis of clusters of genes from microarray studies, including KEGG pathway mapping, GO term analysis, promoter analysis and chromosomal location. Results can be downloaded as PDF or text files.	(Bastien et al. 2004; Joubert and Joubert 2008; Law et al. 2008)

Table 1.5. A list of online *Plasmodium* genome databases and resources. They are defined here as being dedicated whole-genome scale information resources that provide searchable interfaces and/or tools to analyse the genes and gene products and collate information about them.

A further limitation of the resources listed in Table 1.5 is the primitiveness of their graphical displays. For example, PlasmoDB visualisation runs to a basic

genome browser that is able to display genes in the context of a small region of the chromosome (in theory, up to 4 mb; in practice up to 1 mb) and a set of small expression profile graphs.

Indeed, a recent development in the field of *Plasmodium* bioinformatics, the MalPort web server (MalPort; http://malport.bi.up.ac.za:7070/), represents a stride in the direction to addressing these short-comings. By providing tools (MADIBA) for the analysis of gene clusters from microarray experiments (Law et al. 2008), MalPort emphasises the need to provide biologists with new tools allowing them to explore trends across gene groups. Moreover, MalPort introduces the use of basic visualisation tools in analysis, such as providing KEGG pathway maps and displaying genomic location of clustered genes.

1.6.2 Related tools for organisation and analysis of generic or other organism-specific functional genomic data

In the wider-field of functional genomic data representation and analysis many tools, both generic and organism-specific, have been developed to facilitate biologists and bioinformaticians to explore the data.

Tools for mining functional genomic data can be divided into three broad categories: data management, annotation/analysis, and exploration/discovery. Table 1.6 lists examples of tools that offer integrated solutions for data management. Table 1.7 lists examples of tools that are designed to facilitate data annotation and/or analysis. Finally, Table 1.8 lists examples of tools that help users to explore data using graphical displays.

Visualisation tools for functional genomics data are proven to help biologists and bioinformaticians better understand the data and make sense of complex patterns

and relationships. A large number of tools offering advanced visualisation capabilities for various types of functional genomics data are now available. Many of them focus on the display of a single dataset, principally for the purpose of helping experimentalists explore and annotate the data they have generated, such as from microarray experiments. Increasingly, tools are emerging that allow users to visualise multiple data-types concurrently and to overlay results from two or more experiments.

The development of such tools is driven primarily by the vision of companies seeking to provide the answer to a laboratory's ever-evolving data management and analysis requirements. Hence, the products are typically designed for use with the vast quantity of data generated by medically-relevant human and mouse research. The pressing need to develop a centralised system for the organisation and mining of varied information about *Plasmodium* genes has become strikingly apparent (Birkholtz et al. 2006).

Here I present a new development designed to fill this gap in the market by way of an innovative software tool providing graphical interfaces to various *P. falciparum* functional genomic data.

Tool	Web address	Publication	Summary	Features	Data-Types	Availability
Flymine/ Intermine	http://www.flymine.org http://www.intermine.org	(Lyne et al. 2007)	Integrated database for genomic, protein and expression data for *Drosophila*, *Anopheles* and *Caenorhabditis elegans*	Powerful query builder interface; ask questions involving multiple genes; template queries provided; integrates many different data-types; import/export gene lists.	Comparative genomics, proteins structures, interactions, GO, expression, transcriptional regulation, pathways, literature.	Freely available to use online.
Catalyzer by Axiope	http://www.axiope.com/oidex.html	(Goddard et al. 2003)	Flexible system for managing and sharing laboratory data. Designed for non-IT experts.	Several plug-ins available for specific biomedical data-types; proteomics and genomics plug-ins planned; software development kit available to developers.	Almost any type of data that can be catalogued.	Limited trial version is free; purchase full version.

Table 1.6. Examples of software for managing the storage and searching of integrated biological data.

Tool	Web address	Publication	Summary	Features	Data-Types	Availability
Ensembl	http://www.ensembl.org	(Flicek et al. 2008)	Provides an automated pipeline and set of tools for genome annotation and visualisation.	All gene predictions based on experimental evidence; can be adapted for many organisms; upload of custom datasets possible; chromosome browser with synteny maps; detailed sequence views.	All gene types, SNPs, exons, comparative genomics, ESTs, alternative accession numbers, transcripts, protein features including GO and InterPro.	All data and software free to use and download.

UCSC Genome Browser	http://genome.ucsc.edu	(Kent et al. 2002)	A set of tools for visualisation and analysis of genomic data. Different datasets are displayed as multiple 'tracks' next to the genome sequence to facilitate comparison.	Search for groups of genes based on relationships, e.g. homology, expression profile etc; map uploaded sequences and genome-wide datasets (such as SNPs and linkage studies) to a genome; download database Tables; summary and links to protein properties, metabolic pathways and structures.	Chromosomes, genes, expression data, ESTs, SNPs, potentially customisable to include any other type of genomic-scale data, proteins and protein structures.	Freely available to use online. License required for commercial use of source code and execuTables; free for everyone else.
TableView	http://www.ccgb.umn.edu/software/java/apps/TableView/	(Johnson et al. 2003)	A general purpose application for exploration and visualisation of tabular data.	Offers multiple views of the data, including histograms, hierarchical clusters, 2-D and 3-D scatter plots. Selected subsets of data points are carried forward between viewers.	Can be used with a variety of biological data including ESTs, SAGE, microarray and annotation data.	Application and source code are freely available to download.
GenePilot by TG Services, Inc.	http://www.genepilot.com	N/A	Analysis suite for microarray data, providing a variety of tools including clustering algorithms and heat maps.	Results link to a customisable GO display for a group of genes; export results as images; light-weight - runs on computers with low memory.	Microarrays including expression data and ESTs, GO, gene annotations and link-outs to other databases.	Free to academics; fee for commercial use. Can be customised for a fee.
Bioverse	http://bioverse.compbio.washington.edu/	(McDermott et al. 2005)	A sequence annotation server with web-interface for representing biological components and relationships between them.	Users can compile and annotate lists of proteins; Integrator – a viewer for large interaction networks, to which users can upload gene expression data etc.; offers confidence scores for all predicted annotation.	Sequences, secondary and tertiary experimental and predicted protein structures, functional motifs, interactions, homologs.	Freely available to use online.

MultiExperiment Viewer by TM4	http://www.tm4.org/mev.html	(Saeed et al. 2006)	Versatile microarray analysis tool offering a wide range of algorithms for clustering, visualisation, classification and statistical analysis.	Clusters of genes can be tracked through multiple analyses; modules for viewing metabolic pathways and genome/chromosome maps, with expression data overlaid, are in development; links to public databases are also in development.	Several microarray file formats accepted.	Free to download.
DecisionSite by Spotfire	http://spotfire.tibco.com/products/decisionsite.cfm	N/A	A suite of software tools providing general and specific solutions for interactive, visual data analysis and exploration.	DecisionSite for Functional Genomics – for analysis and visualisation of gene expression data, clustering, isolation of genes of interest, hypothesis testing, access to data from user files, databases and web sources, developer package to add new algorithms. DecisionSite for Microarray Analysis – offers more advanced analysis, assessment and statistics for processing raw data and complex experiments.	All types of microarrays, GO, link-outs to online databases, such as PubMed.	Available to buy.

Table 1.7. Examples of software that use interactive graphical displays for annotation or analysis of genomic or functional genomic data.

Tool	Web address	Publication	Summary	Features	Data-Types	Availability
Pathway Tools	http://bioinformatics.ai.sri.com/ptools/	(Karp et al. 2002)	A set of systems biology tools for supporting development of organism-specific database of metabolic pathways, genome browser, transcriptional regulatory networks, functional genomics data integration and comparative analysis.	Software includes: PathoLogic for creating a Pathway/Genome Database (PGDB) of predicted metabolic pathways for an organism; Pathway/Genome Navigator for queries, visualisation and analysis of a PGDB; Pathway/Genome Editors for editing PGDBs; Omics Viewer for displaying single experiment or time-series quantitative (e.g. expression) data painted onto a cellular overview or genomic map.	Genes, proteins, pathways, transcriptional control, gene or protein expression data, metabolomics data.	Free to academics; fee for commercial use.
Yeast Exploration Tool Integrator	http://www.yetibio.com/	(Orton et al. 2004)	Software tool for the visualisation and analysis of yeast functional genomics data. Integrates a range of genomic, interaction, functional annotation and expression data.	Set of linked viewers for different data-types, including a genome/chromosome viewer, expression data, protein interactions and annotation Tables; sets of genes can be selected and carried across viewers; advanced visualisation including overlaying expression data onto genomic location and networks.	Chromosomes, genes, GO, protein interactions, hierarchically clustered expression data, phenotypes.	Free to use online after registration. Application and database can be downloaded under license.

GenMAPP	http://www.genmapp.org	(Salomonis et al. 2007)	Pathway-based approach for visualising and analysing expression data. Includes a database and multiple data viewers, including pathways, clusters of co-expressed genes and Tables.	Visualisation of integrated data-types; mRNA and/or protein expression (including time-course), SNP or splicing data mapped onto a pathway; create custom species database and pathways; export maps as HTML; compare pathways between species.	Gene and protein expression data (including time-series), SNPs, sequence variants, pathways, GO.	Free to use after registration. Source code is available under Apache open source license.
Osprey	http://biodata.mshri.on.ca/osprey/servlet/Index	(Breitkreutz et al. 2003)	A software tool for visualisation of complex interaction networks. Builds networks using published interactions and GO data maintained by The GRID.	Networks can be coloured by gene function (GO) and/or experimental system; Advanced filtering and layout options; facilitates comparison between datasets; interaction datasets in several standard file formats can be loaded in; networks can be saved and exported in various formats.	Protein and gene interaction networks, GO and annotation data.	Free to download after registration for not-for-profit organisations. Limited version online.
GeneXplorer	http://www.gmod.org/wiki/index.php/GeneXplorer	(Rees et al. 2004)	An application to generate interactive microarray dataset visualisation and analysis that can be viewed in a web browser.	The resulting web application has many of the features of similar standalone software; gene annotation and link-outs are fully customisable; create a list of genes with similar expression patterns.	Accepts clustered microarray data in the Clustered Data Table (CDT) format. Output is HTML and JavaScript.	Executables and source code are freely available to download.

DAVID	http://david.abcc.ncifcrf.gov/	(Dennis et al. 2003)	Database for Annotation, Visualisation and Integrated Discovery. Set of tools for discovering biological meaning within large lists of genes.	Retrieve functional annotation for lists of genes and cluster them by function; discover enriched terms (e.g. GO); map genes to pathways and disease; search for homologs and literature; convert gene IDs to other accession numbers;	Detailed gene annotation, links to source databases and literature, GO, protein domains.	Web-based tools are free to use. Download is free to non-profit users, license required for commercial use.
PathwayExplorer	https://pathwayexplorer.genome.tugraz.at/	(Mlecnik et al. 2005)	Web service for visualising expression data mapped onto biological pathways.	Pathways available from KEGG, BioCarta and GenMAPP; export maps as graphics; users upload their own expression data; filtering of expression data; expression profiles for genes can be displayed.	Available regulatory, metabolic and cellular pathways, expression data, links to a few other databases.	Freely available to use online or as a stand-alone version.
SHARKview	http://bioinformatics.leeds.ac.uk/shark/	(Hyland et al. 2006)	Web-based visualisation tool for exploring and comparing metabolic data from different organisms. Front-end to the metaSHARK package for reconstruction of metabolic networks from unannotated genomes.	Gene expression data can be uploaded from user files; metabolic pathways can be customised and constructed from user-uploaded data; link-outs to databases such as KEGG and PRIAM are provided; users may save data and pathways in their own accounts.	Metabolic data produced by the SHARKhunt tool, outputs from other software such as PRIAM and Pathway Tools, gene expression data in text format.	Freely available to use online. Register for free account to access advanced features.

Table 1.8. Examples of software tools that facilitate exploration by providing visualisation of integrated functional genomics data.

1.7 Motivation

The value of data produced in large-scale functional genomic studies can only be harnessed by the research community if there are effective tools to explore and mine them. Graphical display greatly enhances the information that can be gleaned by biologists. Integration of different data-types affords users the advantage of being able to examine multiple lines of evidence about a gene's function.

In the field of malaria research very little is known about the roles of the parasite's genes and their protein products. Resources for research have been relatively limited due to malaria drugs and vaccines being non-profitable for pharmaceutical companies. The malaria research community would benefit greatly from a publicly-available software tool for visualisation of integrated functional genomics data. When research for this PhD thesis began in late 2004 the genome sequence of the malaria parasite had been available for two years and the first large-scale functional genomics studies had just been published. During the course of the research, several new studies have been performed, which provide data about many of the parasite's genes that were previously uncharacterised. The task now is presenting the data to biologists in a way that allows them to extract meaningful information about their proteins of interest and to view the results in the context of all the other data.

The importance of a central resource for integrating and mining the malaria molecular, functional and pharmacological data was emphasised in a recent paper by Birkholtz et al. (2006), in which the authors called for the use of modern computational resources to organise the data from functional genomic experiments in

a useful, versatile way to enhance the identification and characterisation of novel targets (Birkholtz et al. 2006).

1.8 Aims and objectives of thesis

In this thesis I introduce the Malaria Genome Exploration Tool (MaGnET), a novel software tool for integrated visualisation of functional genomic data pertaining to *P. falciparum* and related organisms. Chapter 2 describes the design philosophy and the specific aims and objectives of the software. Chapter 3 lists the publicly-available datasets that are stored in the MaGnET database, any processing applied to the data and the structure of the database. Chapter 4 presents the implementation and important features of the MaGnET visualisation program, accompanied by helpful screenshots. Additionally, it discusses MaGnET's advantages over other tools in the field and suggests directions for useful future expansion. Chapter 5 demonstrate how MaGnET can reproduce the results of several recent studies into *P. falciparum* gene function. Chapter 6 reports on analyses performed using MaGnET that has lead to the development of novel hypotheses about gene function, which are testable in the lab. The final Chapter draws conclusions about the work presented in this thesis and puts it in context with other work in the field.

2. SYSTEM DESIGN

Overview

Chapter 1 described the current status of *Plasmodium* functional genomic research and the motivation for developing a new tool to enable biologists to explore the data. In this chapter, specific design proposals and targets for the tool will be described. The chapter starts with definitions of key concepts and layout of the overall software aims. A detailed diagram and description of the system design follow. Identifiable aims for data inclusion, user interface design and software usage will be outlined. Finally, the specific shortcomings of other resources that the tool aims to address will be discussed.

2.1 Definitions

The following section introduces definitions for three concepts central to this thesis.

'Visualisation'

Scientific visualisation is the use of computer graphics to present any kind of scientific data and is particularly helpful when viewing large quantities of data. Visualisation aids reasoning, perception and hypothesis formation. Graphical displays can be anything from simple two-variable graphs to virtual representations of biological entities.

'Integration'

In a bioinformatic sense, integration refers to bringing different biological data together within a computational resource. Data may be displayed side-by-side or one type of data mapped onto another, in order to facilitate their comparison. Integrating disparate data offers challenges for visualisation, and it is important to ensure that the data is displayed in a biologically meaningful manner.

'Exploration'

In scientific research, exploration means investigating the unknown to gain an initial understanding. Exploration often refers to a careful, systematic search leading to new a discovery.

2.2 Software aims

The overall aims for the new software tool are outlined here. The principal aim is to provide malaria biologists with tools for visualisation of various important *Plasmodium* functional genomic datasets. These data include recent genomic, proteomic and transcriptomic datasets, as well as certain selected predictions about protein function, described in Chapter 1. The viewers for different kinds of data need to be integrated and permit users to carry selections between them. To facilitate user interaction with the software they must be able to easily select genes of interest and to modify their selection. The program will allow users to take a novel approach to hypothesis generation by encouraging them to explore available data through visualisation. Ultimately, it should aid users in discovering hitherto unrecognised relationships between genes and to make informed predictions about gene function.

To achieve this, the software tool needs to incorporate a database in which to store the required datasets and a program for rendering the graphical display of data.

2.3 The MaGnET system

Figure 2.1 presents the MaGnET system connectivity map. It is based around a MySQL database (MySQL AB, Uppsala, Sweden; http://www.mysql.com/) to hold the data and a Java program (Sun Microsystems, Santa Clara, CA, USA; http://www.java.com/) for visualisation. Section 3.1 includes a detailed description of the database design and Sections 4.2 and 4.3 describe the design of the Java program and user interface. The data are read into the database by several programs, each written for the specific purpose of reading and processing a particular dataset. Section 3.2 lists the datasets and their sources, and Section 3.3 describes the process by which each dataset is processed to extract the relevant information and prepare it for entry into the database.

The visualisation program is based around four integrated data viewers. The Genome Viewer is for visualising chromosome and gene organisation at the whole genome level, at the level of an individual chromosome and at the level of individual gene intron/exon organisation. The Protein-Protein Interaction Viewer presents a novel method for viewing *Plasmodium* protein interaction networks. The Expression Data Viewer is for visualisation of time-series expression data (mRNA or protein levels recorded over various stages of the life cycle). The Data Analysis Viewer allows querying of the database and displays results and various data about the genes in the form of a table. All the sections are linked and groups of genes selected in one viewer can be easily transported to the other viewers.

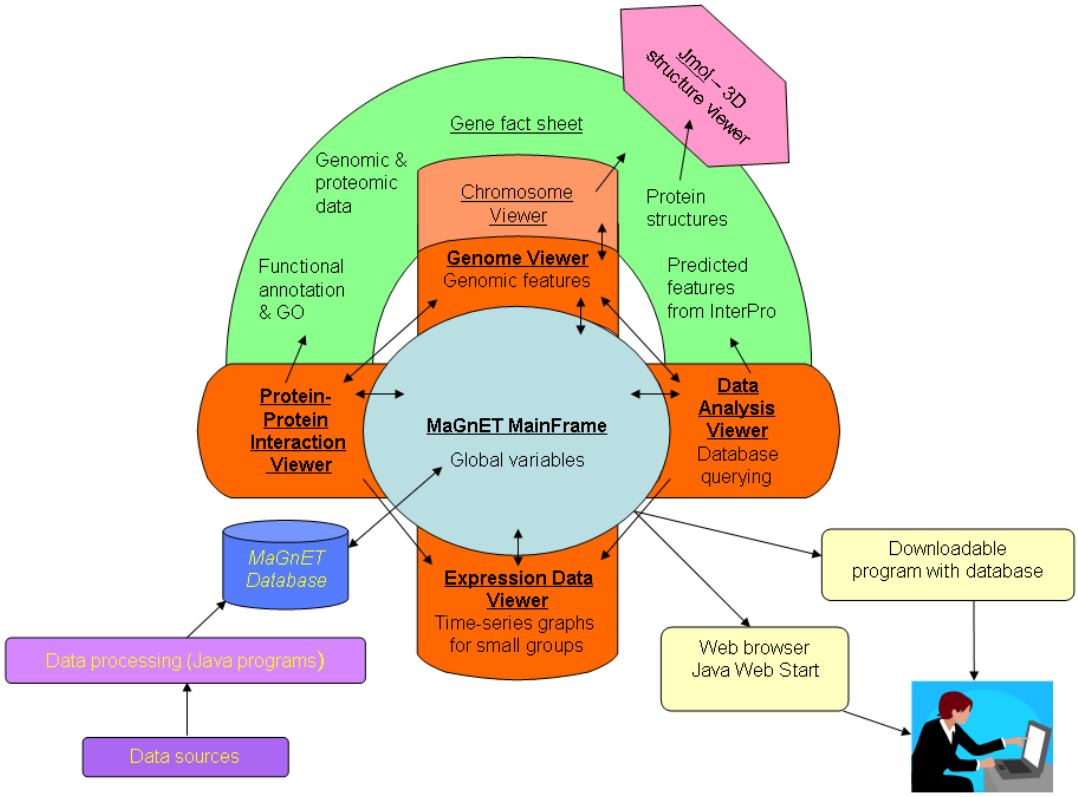


Figure 2.1. The MaGnET connectivity map. Data is read in and stored in a MySQL database and visualised with a Java program. The MainFrame class stores globally-needed information and controls the rest of the program. There are four viewers for visualising genomic features, protein-protein interactions, protein and mRNA expression data and the results of database queries. Gene fact sheets provide a summary of all the data available about a particular gene and link to the Jmol structure viewer for visualising 3D protein structures. The program and database can be used online through a web browser as a Java applet, or as a downloadable application with Java Web Start technology. Alternatively, users may request their own copy of the program and database that can be installed and used locally.

2.4 Objectives for data inclusion

The data included within MaGnET should all come from publicly-available sources, in order that MaGnET itself will be completely free for use. The data

should either come from well-received published studies or from widely-used annotation tools, and be endorsed by the major *Plasmodium* data resources, such as PlasmoDB (Bahl et al. 2003) and GeneDB (Hertz-Fowler et al. 2004). This should ensure that the experimental and predicted data included within MaGnET is of a high-standard. The careful inclusion of selected predicted annotations can provide additional helpful clues towards possible function for genes annotated as "hypothetical proteins" [approximately two-thirds of the predicted proteome following genome sequencing (Gardner et al. 2002)].

Datasets for incorporation in MaGnET should be carefully selected so that they are of relevance for research into gene function. Genome resources often provide every bit of possible information about a gene, which is useful for reference purposes, but some of it is not necessary for formulating new hypotheses about probable functions. Moreover, a specific aim of MaGnET is to minimise the feeling of 'data overload' often felt by biologists when using these kinds of resources. The inclusion of only selected datasets should achieve this.

The data also needs to be a format that can easily be read into the database without too much necessary processing. Downloadable text files in standard formats are best. This will help to ensure that future releases of data files remain compatible and update programs can easily be deployed to update the database.

2.5 Objectives for database design

The software used for the MaGnET database needs to be robust, reliable and able to perform fast searches and data retrieval. It also needs to be able to cope with a variety of data-types. The database design must organise the data intuitively,

without repetition, and needs to be flexible and extendable to allow future expansion. The software program for data visualisation needs to be able to communicate effectively with the database.

The database needs to be in a format that can be freely distributed, so that in future users could download and install a local copy of the database, into which they could add their own datasets.

2.6 Objectives for visualisation program design

2.6.1 Technical requirements

The programming language chosen for developing the MaGnET visualisation program needs to be supported by all other necessary technologies, including the database system, internet applications, such as web browsers, and all major operating systems. It should be well-documented, actively developed and widely used, in order to ensure there is a strong base for support. The latter is essential for reliable, efficient program development because programming language user and developer community websites provide helpful coding examples and advice on trouble-shooting for common problems.

2.6.2 User interface

The visualisation program user interface will provide graphical displays for a range of important functional genomic datasets. The interface will include several dedicated 'data viewers' to encourage users to explore individual datasets.

Moreover, all data viewers will be fully integrated in order that users can move easily between viewers and can carry gene selections forward between viewers.

MaGnET aims to provide users with novel features not currently implemented in available *Plasmodium* data resources. For example, while browsing, users will be encouraged to select groups of genes based on common properties. Relationships between genes and shared properties of the group can then be explored in different contexts by enabling users to maintain their selections across datasets. Furthermore, the user will be able to change their selection at any point, adding and removing genes as they choose, unlike in other tools that allow selection of pre-calculated gene sets.

Furthermore, MaGnET will advance on currently available graphical interfaces by integrating different types of functional genomic data in a single display. This will bring an extra dimension functional genomic data visualisation that will allow users to easily explore relationships between data-types. For instance, visualisation of gene expression data mapped onto the genomic location could lead to discovery of co-regulated genes in close proximity.

The visualisation program must implement a search facility to enable the extraction of textual annotation data from the MaGnET database. This data should be fully searchable via keyword, protein name or gene identifier. Users of functional genomic data resources will always want to search for their particular proteins of interest by name or identifier; initially to test the software out using proteins with which they are familiar and, later, to access data about other proteins they have become interested in.

While exploring, users need to be able to quickly access summaries of relevant functional and structural annotation about individual proteins. MaGnET aims to facilitate this by providing a summary page about each gene that can be accessed at any time.

2.7 Overall objectives for MaGnET

2.7.1 Usability

The resulting software tool needs to be entirely platform-independent, so that the software can easily work on any computer. It needs to work 'out of the box' so that users do not have to go through a complicated set-up procedure to use the tool. It needs to be freely accessible to researchers in academic and government institutions, which is where the majority of fundamental research into malaria happens. To ensure it reaches the people most in need of free software tools for malaria research (developing countries in particular), it needs to be accessible on the internet world-wide and should not require special resources to work.

The user-interface needs to be simple enough for first-time users to quickly learn how to interact with the main features of the tool, and they need to know where to look to find instructions for using less obvious features. Hence, a detailed set of help pages and simple-to-follow tutorials will be important for users to get the maximum advantage out of the program. However, help pages are not usually the first port of call for first-time users, so the program does need to be laid out intuitively. Also, the target audience are mainly laboratory-based biologists, rather than experienced bioinformaticians, so to use the tool's features cannot require any

prior knowledge of bioinformatic techniques. Having said that, bioinformaticians may find some aspects of MaGnET useful, such as advanced functionality for expression data exploration, that are not available in other *Plasmodium* tools online.

Furthermore, the tool needs to make it clear to users what kind of data they are viewing and where it came from, so they can make an informed judgement on how much emphasis they wish to place on the various sources of information. There should be links to the original data and to other helpful sources of information in relevant online resources.

2.7.2 Outcome of MaGnET usage

The ultimate aim of exploratory analyses using MaGnET is the generation of novel hypotheses about gene function, which can then be tested in the laboratory. In order to formulate sound hypotheses users should be encouraged to investigate all lines of evidence available by browsing different aspects of gene and protein function using MaGnET's data viewers. It is anticipated that MaGnET will fill a niche for a tool to kick-start new investigations into gene function by providing the opportunity to explore data that is otherwise quite inaccessible to malaria researchers. MaGnET will hopefully become the starting point for expanding into other detailed analyses, which potentially could be either laboratory-based or involve further computational investigation.

2.8 Specific limitations of related tools that MaGnET aims to address

This section will expand upon the points introduced in Section 1.6 to set out how MaGnET aims to address specific short-comings of existing resources for functional genomic data organisation and display.

The single gene focus of current databases prevents the user from easily comparing annotation of multiple genes. This is important for a number of reasons, but it is particularly advantageous for assigning potential annotation to previously uncharacterised genes. The concept of 'guilt by association' has been well established in the field of genome annotation, where it holds that an uncharacterised gene that shares features in common with a group of genes of known function is likely to be associated with the same biological process or functional goal (Ettwiller and Paten 2004).

Conventional tools rarely encourage users to browse datasets because they often force the user to start with the end in mind. Many of the commonly used databases provide a search facility that requires the user to know what they are looking for, but do not provide a helpful interface to assist users with exploring the data to make discoveries that could lead to testable hypotheses. This kind of approach could be especially useful in the quest for new drug and vaccine targets.

Existing databases tend to present all the data that is available about a particular gene, including predicted information that is generated by the many automated tools for genome annotation (which often infer function based on sequence similarity as discussed in Chapter 1). The problem is that predicted information is not always clearly distinguishable to biologists. Another disadvantage

faced by biologists is that when using such genome-scale databases they sometimes experience a sense of 'data overload' that makes it difficult to extract the relevant data of interest to them. MaGnET aims to deal with the first of these issues by employing quality control criteria for the data included and to clearly mark any predicted annotation as such. To address the latter concern MaGnET will include only the most relevant data for the purposes of gathering clues to potential gene function. Therefore, MaGnET is not aimed at being a replacement for existing resources, but rather to complement them by providing a means for users to explore the major functional genomic datasets and generate hypotheses that they will be able to test using other bioinformatic or biochemical techniques.

Very few of the publicly available tools have implemented extensive visualisation facilities for *Plasmodium* functional genomic data. Where satisfactory visualisation does exist, it is usually limited to one type of data; for example, the Malaria Metabolic Pathways Database provides helpful diagrams of parasite pathways (Ginsburg 2006). Several generic tools that provide visualisation are commercially available and/or require a prohibitively complicated installation procedure. Therefore, the work described in this thesis aims to bridge an evident gap for a tool providing integrated visualisation of *Plasmodium* functional genomic datasets that is both lightweight and publicly accessible.

3. DATA AND DATA PROCESSING

Overview

MaGnET integrates a variety of publicly available functional genomic datasets for *P. falciparum*. The sequencing consortium typically provides downloadable files containing the results of their genome sequencing, automated and manual annotation projects available via the Wellcome Trust Sanger Institute *P. falciparum* Genome Project web pages (http://www.sanger.ac.uk/Projects/P_falciparum/) or the *Plasmodium* Genome Resource (PlasmoDB; http://www.plasmodb.org/plasmo/). Other datasets are provided as publication supplementary material; the data may be provided in a range of raw and processed formats (such as averages of multiple replicates or normalised against control data).

In this chapter, the design and structure of the MaGnET database and the datasets used to populate it are described, along with details of any processing that was performed prior to storing the data.

3.1 Database development

All data used by the MaGnET visualisation program needed to be stored in a dedicated database so that it was readily and quickly accessible. A relational database model was chosen as the most appropriate due to its highly structured nature of storing data within tables. The system of storing records as rows (tuples) and data fields as columns makes it possible to easily compare records and to combine useful data from different tables. The Structured Query Language (SQL)

for relational databases is a powerful, widely-used language that has been incorporated by all the major relational database management systems (RDBMS).

The chosen database management system needed to be portable, expandable, robust, secure, free to use, supported by many platforms and easy to install, with good software and technical support behind it. It also needed to be simple and fast to store data in and read data from, with the facility to design and manage automated queries.

3.1.1 The MySQL database management system

MySQL (MySQL AB, Uppsala, Sweden; http://www.mysql.com/) was the RDBMS of choice since it is one of the most widely used databases around the world. MySQL is fast, reliable, practical and free to use under the Gnu General Public Licence (GPL), as well as meeting all the other aforementioned criteria. MySQL offers extensive application development support, by providing drivers and connectors for major programming languages, such as Java.

The MaGnET database was originally developed using MySQL version 3.23 on an alpha server running Linux hosted by the School of Biology, University of Edinburgh. The development version of the database was later migrated to MySQL version 5.0 on a machine running Windows XP Service Pack 2. The publicly accessible version of the database is currently stored on a server at the University of California, Santa Cruz running MySQL 5.0.

3.1.2 The MaGnET database

The MySQL database has been designed to hold the data with as little redundancy as possible. The tables are easily expandable, so that in the future it will

be possible to add annotation for another genome to the existing database without having to change the database structure. Rows must have a unique, identifying 'primary key'; in many cases unique gene identifiers assigned to genes by the sequencing consortium are used as primary keys.

The MaGnET database holds about 60 megabytes (MB) of data, contained in more than 165,000 rows across 19 tables. When the genomic data for further genomes are added the database size will increase significantly; however, whether the size will remain correlated with the number of genomes depends on how the technology for large-scale functional genomic studies, such as expression, interaction and structural analysis, can keep up with the pace of high-throughput genome sequencing.

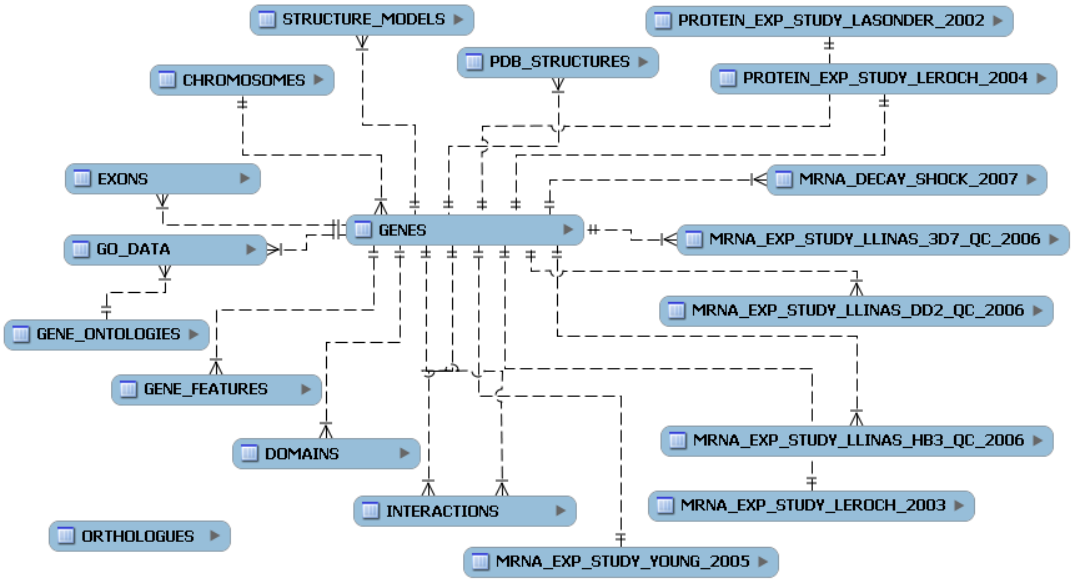


Figure 3.1. Entity relationship (ER) diagram depicting relationships between tables in the MaGnET database. Table columns are described in Appendix A. ER diagram created using MySQL Workbench version 5.0.23 OSS Community Edition (MySQL AB, Uppsala, Sweden; http://www.mysql.com/).

Figure 3.1 is an entity relationship (ER) diagram of the MaGnET database, which shows how the tables are connected. Most of the tables reference a central 'GENES' table, which contains information about individual genes. These include annotation tables, such as 'DOMAINS' and 'GO_DATA' containing data on InterPro predictions and GO assignments, respectively, as well as tables holding individual mRNA and protein expression datasets. Here, the publication first author and year are referenced in the table name, for example the Lasonder et al. 2002 protein expression dataset is stored in table 'PROTEIN_EXP_STUDY_LASONDER_2002'. Full descriptions of the database tables are provided in Appendix A. An SQL file (database 'dump' file) containing all the data tables is provided on the accompanying CD. Section 3.2 describes the datasets included in the database and Section 3.3 describes how and what data are extracted from input files and any processing required prior to adding them to the database.

3.2 Data sets

Table 3.1 summarises the original datasets and sources of the data stored in the MaGnET database. The data come from a variety of different sources, including public databases, genome sequencing and annotation project web pages and journal publication supplementary materials. Appendix B describes in more detail the files that were downloaded, their sources, version numbers and release dates.

Dataset	Source	Data format and description
Chromosome, gene and protein sequences	Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/P_falciparum/) PlasmodDB (http://www.plasmodb.org/plasmo/)	Sequences are provided in FASTA format.
Curated gene annotation	Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/P_falciparum/)	EMBL/GenBank format files containing various aspects of gene annotation provided by the sequencing consortium.
Gene Ontology annotation	Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/P_falciparum/) The Gene Ontology (http://www.geneontology.org)	GO annotation file format. GO term descriptions are downloaded in Open Biomedical Ontologies (OBO) 1.0 file format. See Section 1.4.3.1 for more details.
InterPro predicted sequence features	Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/P_falciparum/)	Tab-delimited text file. See Section 1.4.3.2 for more details.
Ortholog and paralog groupings	Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/P_falciparum/) OrthoMCL (http://orthomcl.cbil.upenn.edu/)	OrthoMCL output format (tab-delimited text file)
Protein-protein interactions	LaCount et al. (2005) A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* 438: 103-107.	Tab-delimited text file listing pairwise protein interactions determined by yeast two-hybrid experiments. The most promiscuous fragments were removed in order to minimise the number of non-specific interactions. See Section 1.5.3 for more details.
Experimentally-solved 3D protein structures	RCSB Protein Data Bank (http://www.rcsb.org/pdb)	Structure coordinates are available as PDB format files and sequences are available in FASTA format.
Comparatively-modelled 3D protein structures	ModBase (http://modbase.compbio.ucsf.edu/)	Structure coordinates are available as PDB format files. See Section 1.5.4.2 for more details about ModBase.

mRNA time-series expression data for several life cycle stages	Llinas et al. (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res* 34: 1166-1173. Young et al. (2005) The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* 143: 67-79. Le Roch et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301: 1503-1508.	The datasets produced by Le Roch et al. and Young et al. represent an approximation to absolute gene expression levels. Data have been normalised across probe sets and for duplicate experiments. See Section 1.5.1.1 for further details about the datasets and array. The Llinas et al. datasets used in this project represent normalised ratios of mRNA abundance, rather than absolute values. For genes with multiple probes, each probe's data are included separately. To maintain data quality, only those probes with data across at least 60% of the time-series are included (recommended by the authors). Data have been normalised to background controls and averaged for duplicate experiments. See Section 1.5.1.2 for further details about the datasets and array.
Protein time-series expression data for several life cycle stages	Le Roch et al. (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res* 14: 2308-2318. Florens et al. (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature* 419: 520-526. Lasonder et al. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* 419: 537-542.	Data represent total spectral counts for individual proteins recorded by mass spectrometric analysis of protein content. Spectral counts can be interpreted as a measure of protein abundance (Florens et al. 2002). Two datasets were used in this project: the Lasonder et al. dataset and the Le Roch et al. dataset, which combines the results of Florens et al. with further data. See Section 1.5.2 for more details about the experiments.

Table 3.1. Datasets used to populate the MaGnET database, with details of sources, file formats and any pre-processing carried out on the data prior to downloading.

3.3 Data extraction and database population

The following sections describe how and what data are extracted from the input files described in Section 3.2. Java programs (and some Perl programs) were written to read in the data and use it to populate the database. These programs can be used to easily update the database whenever a new version of a data file is released. Moreover, during the course of this project there have been several modifications to *P. falciparum* gene models and annotation and the availability of this set of programs ensured the database remained regularly updated. The update programs are listed in Appendix C and the source code and binaries are included on the accompanying CD.

3.3.1 Extracting chromosome data

Files containing chromosome sequences were downloaded, and processed as described in Figure 3.2. The data extracted from the files and further information derived from these are listed in Table 3.2.

Extracted data	Derived information
chromosome number/identifier	chromosome length
species and strain	

Table 3.2. Data extracted and derived from chromosome sequence files. Chromosome nucleotide sequences were not saved due to space limitations in the database.

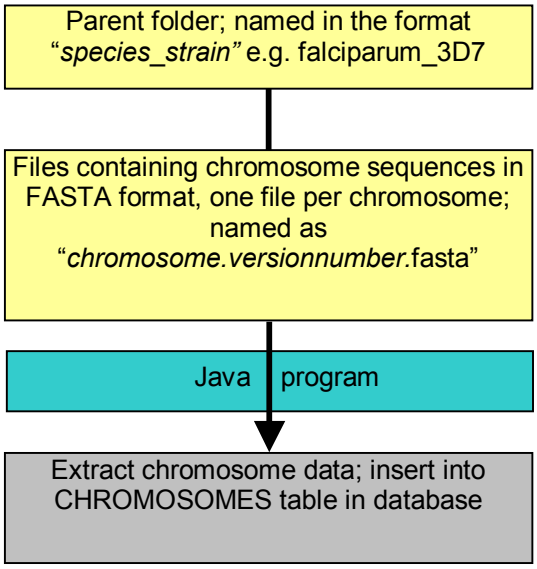


Figure 3.2. Flowchart showing the process of chromosome data extraction.

3.3.2 Extracting gene data

The files containing gene and protein sequences and annotation were downloaded and processed as described in Figure 3.3. The data extracted from the files and derived from these are listed in Table 3.3.

The gene sequence files were read first to retrieve the gene identifiers, chromosome numbers and nucleotide sequences, which were used to create gene entries in the database. The other genomic data and annotation was then retrieved from the EMBL format files (for chromosomes 1-14) and matched to existing entries in the database. The apicoplast and mitochondrial chromosome annotations were available as GenBank format files (very similar to EMBL files); however, there was only limited information available about their genes. Lastly, the protein sequences were extracted and added to the corresponding gene's database record.

The gene identifiers extracted here were used to identify the corresponding gene in all other datasets. Any alternative identifiers by which this gene might be known were read from the EMBL annotation file and stored in the 'alias' field of the gene's record in the database.

Extracted data	Derived information
standard unique gene identifier (gene id)	unique MaGnET identifier (magnet id)
species and strain	length of gene
nucleotide sequence	length of each exon
chromosome	
strand ('w' for Watson, or forward, strand, 'c' for Crick, or reverse, strand)	
location on chromosome: start and end positions	
number and location of exons	
alias (previous/obsolete gene identifiers)	
type (e.g. protein coding, tRNA, pseudogene)	
protein sequence	
product name	
keyword	
detailed annotation provided by the Wellcome Trust Sanger Institute	
signal peptide location and prediction score	
signal anchor location and prediction score	
cleavage site location and prediction score	
standard unique gene identifier (gene id)	

Table 3.3. Data extracted and derived from gene/protein sequence (FASTA) and annotation (EMBL) files.

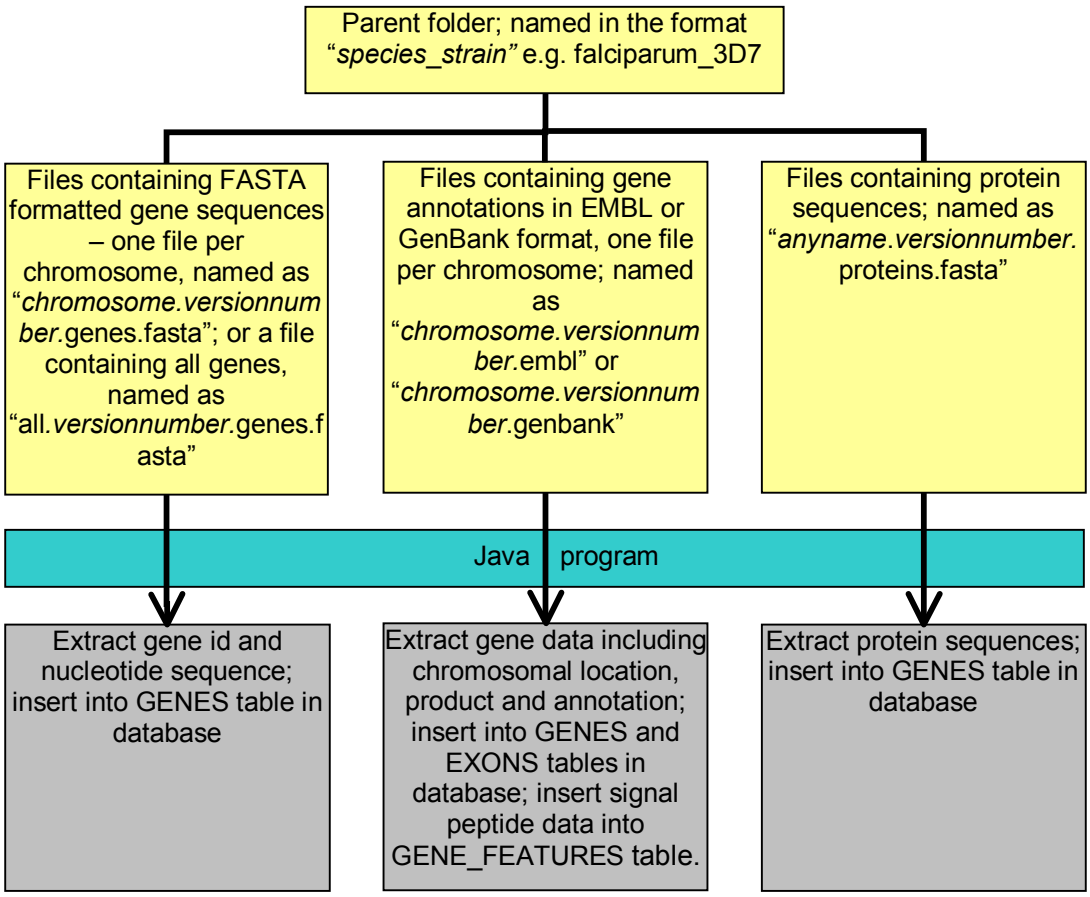


Figure 3.3. Flowchart showing the process of gene data extraction.

3.3.3 Extracting Gene Ontology (GO) annotation

The files containing gene product GO annotations were downloaded and processed as described in Figure 3.4. The data extracted from the files are listed in Table 3.4.

A file containing all GO terms and their full descriptions was downloaded (in OBO format) from the GO website (http://www.geneontology.org), the term names and descriptions were extracted and added to the GENE_ONTOLOGIES table in the

database. The *P. falciparum* GO annotation file (in GeneDB output format) was processed to retrieve the GO identifiers that have been assigned to each gene, if any, and the associated information. Then the GO identifiers were matched to the correct term names in the GENE_ONTOLOGIES table and the *P. falciparum* annotations were added to the GO_DATA table.

Extracted data	**Derived information**
standard gene identifier	a unique annotation identifier
ontology aspect (biological process, molecular function or cellular component	
GO term identifier	
GO term name	
alternative GO term identifier	
description of GO term	
evidence tag (a code indicating how the annotation was made, e.g. ISS – Inferred from Sequence or Structural Similarity)	
evidence with/from (another sequence or annotation which contributed to this annotation, e.g. a similar sequence or interacting protein)	
reference (the database name and accession number, e.g. the publication's PUBMED id)	
date	

Table 3.4. Data extracted and derived from the GO annotation and term description files.

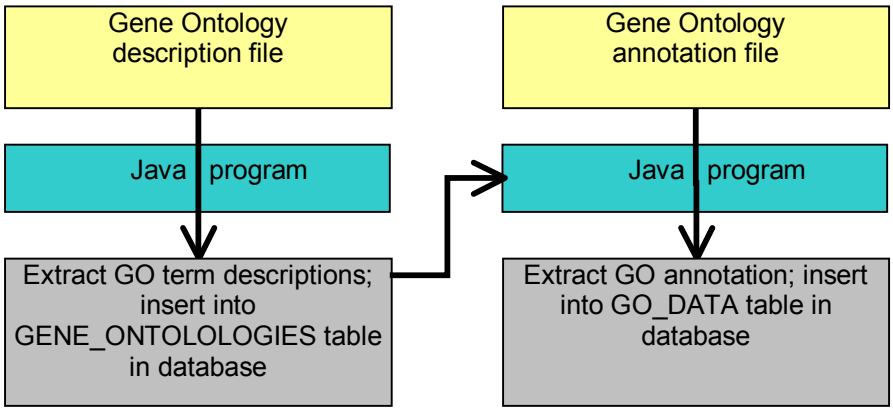


Figure 3.4. Flowchart showing the process of Gene Ontology annotation extraction.

3.3.4 Extracting ortholog and paralog group data

The file containing ortholog and paralog group associations was downloaded and processed as described in Figure 3.5. The data extracted from the file are listed in Table 3.5. This pre-computed cluster file was generated by the genome curators at the WTSI using the program OrthoMCL (Li et al. 2003) and includes five of the currently available and finished *Plasmodium* genome sequences.

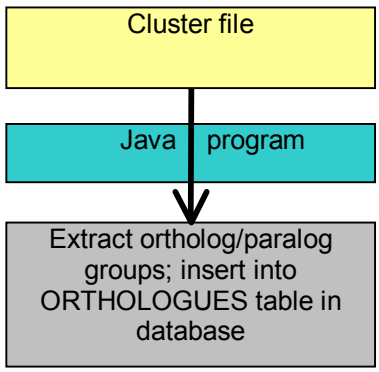


Figure 3.5. Flowchart showing the process of gene ortholog and paralog group extraction.

Extracted data
unique group (cluster) identifier
the number of group members belonging to each of *P. knowlesi*, *P. berghei*, *P. chabaudi*, *P. vivax* and *P. falciparum* 3D7
the gene identifiers of the group members

Table 3.5. Data extracted from the ortholog/paralog cluster file.

3.3.5 Extracting interaction data

The file containing interactions from a large-scale yeast two-hybrid study of protein-protein interactions was downloaded and processed as described in Figure 3.6. The data extracted from the file and derived from these are listed in Table 3.6. The interaction file was read by the Java program and data about the interactions extracted as described above. The program populated the INTERACTIONS table with this data and then used it to calculate the number of unique proteins that each protein in the dataset interacted with.

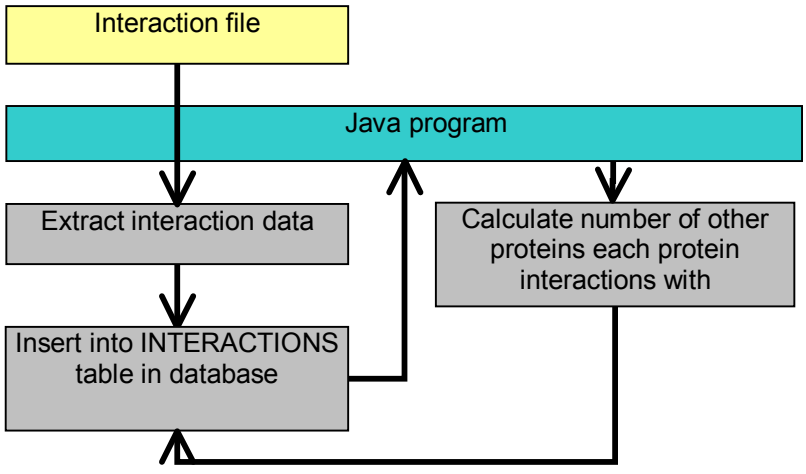


Figure 3.6. Flowchart showing the process of protein-protein interaction data extraction.

Extracted data	Derived information
standard gene identifiers of the bait and prey proteins in the interaction	the number of unique interactions in which the bait protein participates
the number of independent yeast two-hybrid searches in which the interaction was observed	the number of unique interactions in which the prey protein participates
the total number of times the interaction was observed	
the number of prey proteins that interact with the bait protein	
the number of bait proteins that interact with the prey protein	
the type of interaction (e.g. 'self' where two molecules of the same protein interact, or 'reciprocal' where the proteins are observed to interact in either position)	
the name of the study	

Table 3.6. Data extracted from the yeast two-hybrid protein-protein interaction study file.

3.3.6 Extracting protein predicted sequence feature and domain information

The protein predicted sequence feature and domain annotations were downloaded and processed as described in Figure 3.7. The data that are extracted from the file are listed in Table 3.7. The information was read and extracted from the files by a Java program, measured against a cut-off E value (if applicable – not all types of predictions have an associated E value, such as transmembrane and low complexity regions) and inserted into the DOMAINS table in the database.

Extracted data
standard gene identifier
type of domain (e.g. transmembrane, coiled coil, Pfam)
specific domain identifier (e.g. Pfam domain identifier)
description of the domain
start and end positions
expectation (E) value for the domain prediction less than or equal to $1E^{-6}$ (if applicable)
date that the annotation was made
InterPro identifier for this domain (if applicable)
any additional details

Table 3.7. Data extracted about predicted protein sequence features and domains.

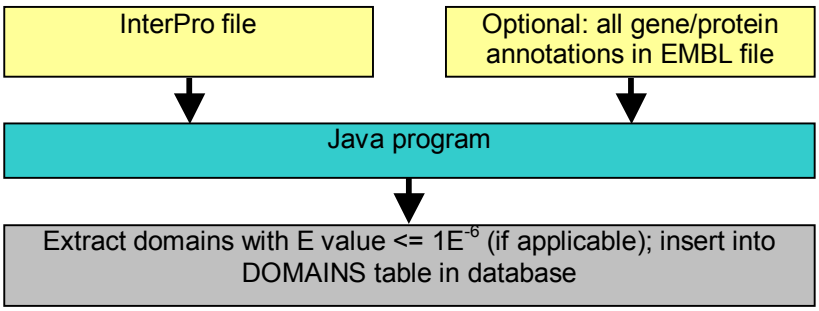


Figure 3.7. Flowchart showing the process of data extraction for predicted protein sequence features and domains. Most annotations were contained in the InterPro file. A recent update to the InterPro file no longer includes transmembrane region predictions. Transmembrane region predictions were retrieved from WTSI genome annotation file (EMBL format).

3.3.7 Retrieving experimentally-solved protein structures

The structures were downloaded as PDB-format coordinate files (one file per solved structure, which can contain multiple protein chains) as described in Figure 3.8. The coordinate files are not currently stored in the MaGnET database due to buffer size and storage space restrictions, so they are stored on the MaGnET server (or in the local file system for local versions of MaGnET) in folders arranged by

Plasmodium species, under a parent folder named 'pdb_structures'. Certain information about the protein chains that have solved structures was extracted and stored in the database. The data extracted from the files and information derived from these are listed in Table 3.8.

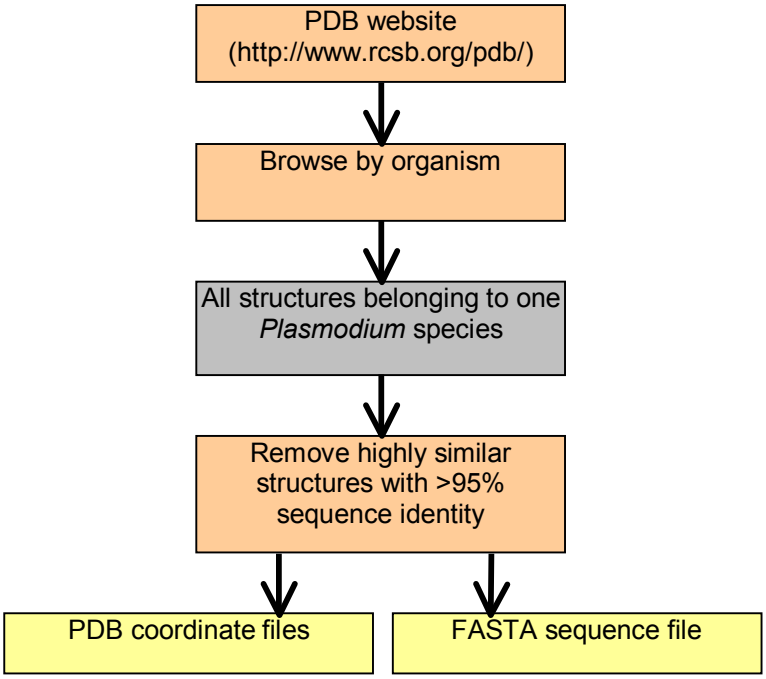


Figure 3.8. Flowchart showing the process by which solved protein structures were extracted from the PDB. A PDB-format coordinate file for each structure and a FASTA file containing all the sequences were downloaded.

The coordinate files were named with their PDB code (e.g. 1N81.pdb). Therefore, in order to utilise the structure files a list containing the PDB code and chain identifier for a particular gene product was required. The database table provided a useful means of storing this information for quick access. Since the gene identifier is rarely included in the coordinate file the structures had to be matched to their corresponding gene identifiers using the process described in Figure 3.9. This

process made use of two programs provided by the National Center for Biotechnology Information (NCBI). The 'FORMATDB' and 'BLASTP' programs are available as part of the BLAST downloadable software package from ftp://ftp.ncbi.nih.gov/blast/. The program FORMATDB creates a local database of sequences against which a search can be run using an appropriate BLAST program, in this case searching for similarity between the solved structure sequences and a database of all protein sequences using BLASTP.

The Java program that processed the BLAST result file assumed that the top hit was the protein whose structure was solved. The information listed in Table 3.8 was then extracted from the FASTA sequence and BLAST result files.

Extracted data	Derived information
PDB code (a unique four letter code identifying the structure in the PDB)	standard gene identifier
chain (the protein chain identifier)	a unique structure identifier
amino acid sequence of the solved region	
starting residue number of the solved region	
ending residue number of the solved region	

Table 3.8. Data extracted and derived from PDB structure files.

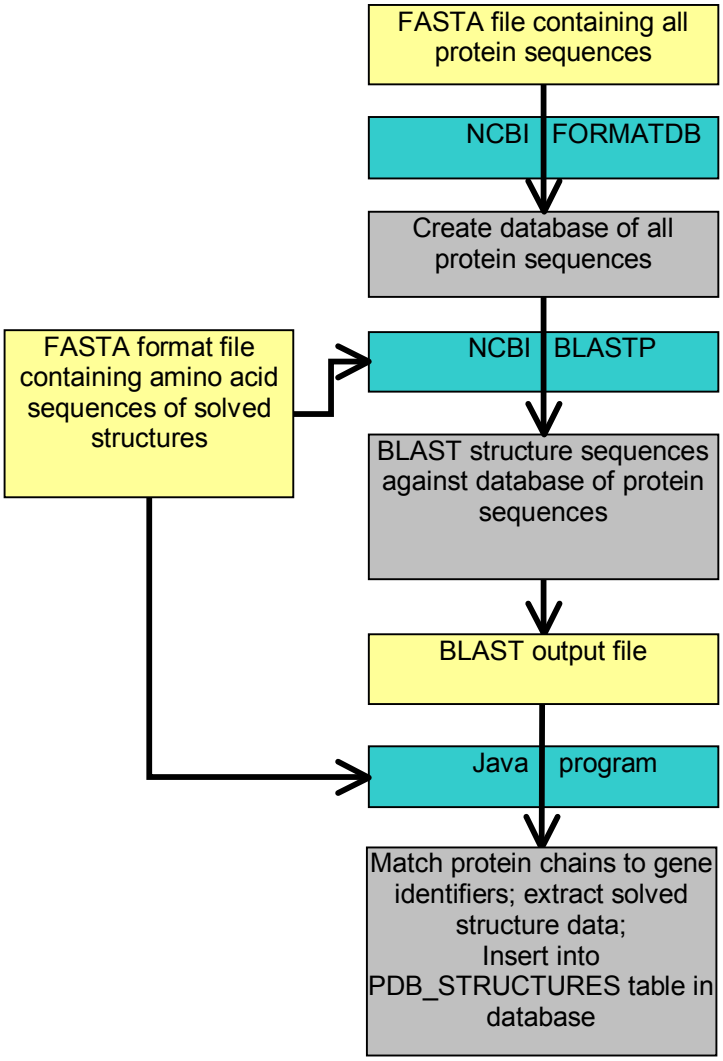


Figure 3.9. Flowchart showing the process of matching solved protein structures with their corresponding gene identifiers and insertion of data about the structure into the MaGnET database.

3.3.8 Retrieving comparatively-modelled protein structures

The modelled structures were downloaded as PDB format coordinate files as described in Figure 3.10. The files were indexed using their standard gene identifiers. The program Wget (Gnu Wget; http://www.gnu.org/software/wget/) automatically downloads the files when provided with a list of URLs (for example, http://salilab.org/modbase/retrieve/modbase/?databaseID=PF14_0433).

Multiple models may be available for an individual gene product and all the models are included within the same file. There are a number of reasons why multiple models could be generated for a protein, including the availability of template structures of varying sizes and representing different structural domains. Another reason why multiple models are often created is due to the availability of templates with differing sequence identities to the query sequence (the modelling process does not filter out templates with a low sequence identity). This leads to a large amount of redundancy in the database, with multiple models representing roughly the same region of protein sequence. It also leads to some low quality models creeping in due to the low sequence identity of target-to-template pairing accepted for modelling. On occasions where a small section of protein sequence has similarity to a small region of a solved structure, a very short model is created. These short models might comprise only a small secondary structural unit such as an alpha helix, which imparts no information about the overall structure of the protein.

Despite the deficiencies of the ModBase database of modelled structures, it does contain a large number of useful models of high quality, which makes it an important resource for biologists. However, due to the large number of redundant, low quality and potentially misleading models, accessing the interesting models becomes awkward and off-putting. In addition, many bench biologists are unlikely to know how to distinguish the useful, high quality models from the noise.

In order to collect only the useful models by removing the low quality models and redundancy from the set, the methodology described in Figure 3.10 was developed. A set of criteria was established which models had to match or exceed in order to be retained in the set of high quality models (Table 3.9).

Criteria for model selection
sequence identity of match to template structure >= 20%
expectation (E) value of match <= 1 x 10^{-6}
sequence length of model >= 45 residues
ModBase model score >= 0.7

Table 3.9. Cut-off criteria for comparative model selection. Models must meet or exceed these criteria to be considered good quality.

The E value is the probability of getting a match between two sequences based on chance. The lower the number the greater the probability that the two sequences are truly homologous. The value 1 x 10^{-6} is the generally accepted minimum cut-off for BLAST searches. The ModBase model score is calculated from statistical potentials to predict how good the model is expected to be (Melo et al. 2002). According to the authors, a model is predicted to be reliable if it has a score greater than 0.7. In this case, the probability of the model having the correct fold is above 95%, for which at least 30% of the C alpha atoms must superpose within 3.5Å. The sequence identity between template and target sequence to produce the correct fold needs to be a minimum of 20% (D Gerloff, personal communication). The minimum sequence length of 45 residues was decided by the size of the smallest known structural domain, an EF hand.

Following filtering out of low quality models, the program BLASTCLUST (NCBI; ftp://ftp.ncbi.nih.gov/blast/) was utilised to gather the structures into groups of high similarity based on having an identical sequence over 90% of their length. As described in the previous section the program FORMATDB was first used to build a database of all modelled sequences. The model sequences were not available to download, so they needed to be calculated from residue information in the

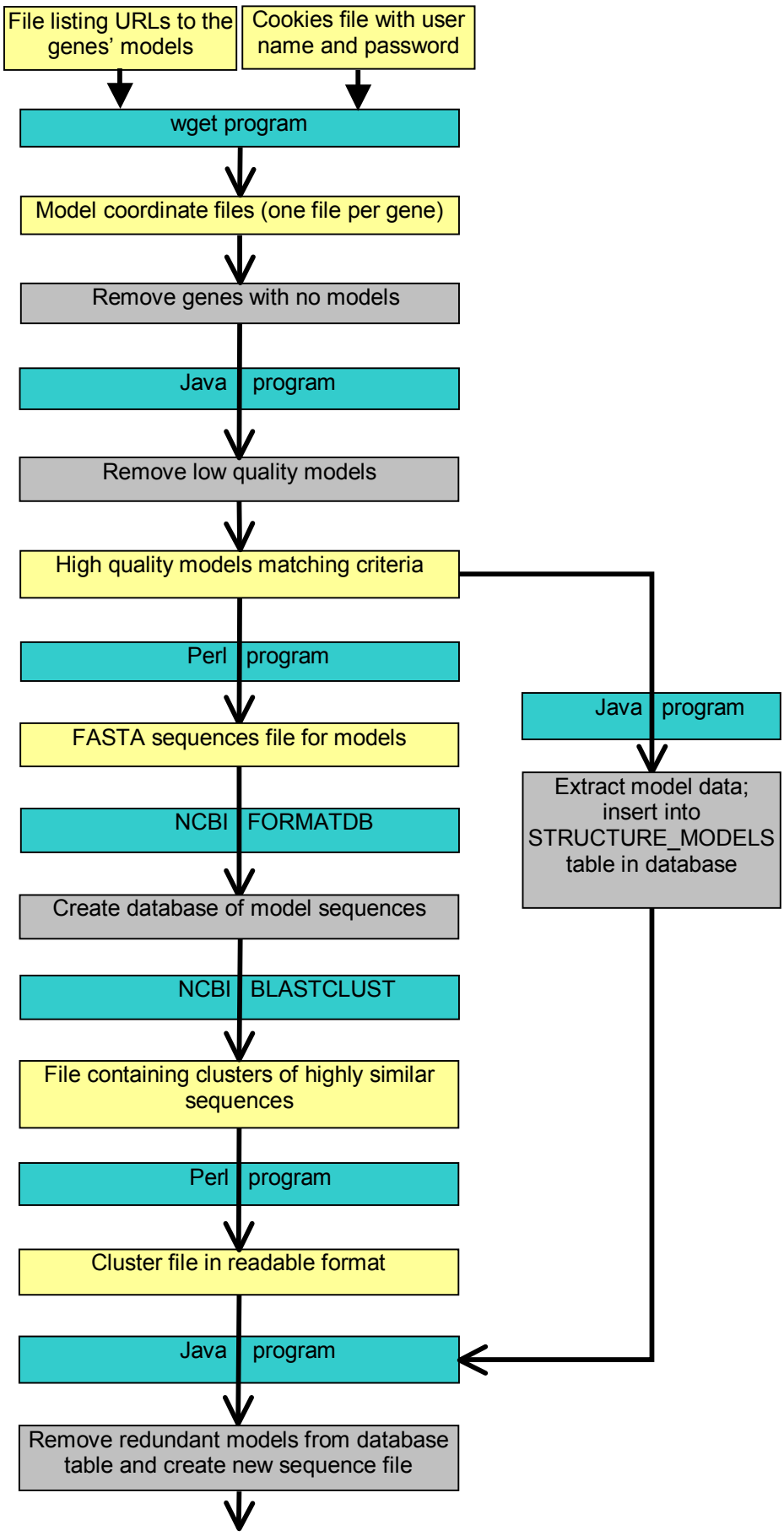
coordinate files. Two rounds of clustering were performed, the first looked for sequences that were identical across greater than 90% of both their lengths and retained the sequence (and hence structure) with greatest sequence identity to template. The second looked for sequences which were sub-parts of longer sequences, by looking for sequences that were identical across greater than 90% of one of their lengths. The shorter sequences were discarded, unless they happened to have sequence identity to template over 5% greater than that of the longer sequence, in which case both sequences were retained. The shorter model is likely to be better due to a significantly higher sequence identity, but the longer model is informative over more of the protein sequence.

As for the experimentally-solved structures, information about the models was stored in a table in the MaGnET database. The data extracted and stored are listed in Table 3.10.

The model coordinate files are stored on the MaGnET server (or local file system for local version), within folders divided according to the species, and inside a parent folder called ‘models’.

Extracted data
a unique model identifier
amino acid sequence of the model
gene identifier of the modelled protein
sequence identity of modelled sequence to template structure
model score
E value of the match between modelled sequence and template
template structure's PDB code
chain identifier of template protein in PDB file
length of the modelled protein
starting residue number of the modelled part of the protein sequence
ending residue number of the modelled part of the protein sequence
starting residue of the template sequence that was used for modelling
ending residue of the template sequence that was used for modelling
date the model was created
the name of the modelling run in which the model was created
any notes about the model

Table 3.10. Data extracted from comparative model structure files.


File listing URLs to the genes' models
Cookies file with user name and password
wget program
Model coordinate files (one file per gene)
Remove genes with no models
Java program
Remove low quality models
High quality models matching criteria
Perl program
FASTA sequences file for models
NCBI FORMATDB
Create database of model sequences
NCBI BLASTCLUST
File containing clusters of highly similar sequences
Perl program
Cluster file in readable format
Java program
Remove redundant models from database table and create new sequence file
Java program
Extract model data; insert into STRUCTURE_MODELS table in database

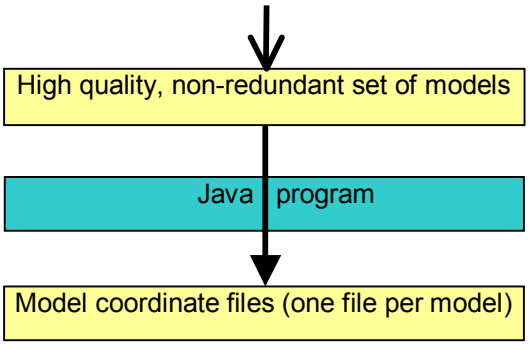


Figure 3.10. Flowchart showing the process of retrieving comparative structure models, filtering out low quality models and removing a large number of redundant models to create a high quality, non-redundant set of representative models.

3.3.9 Extracting expression data

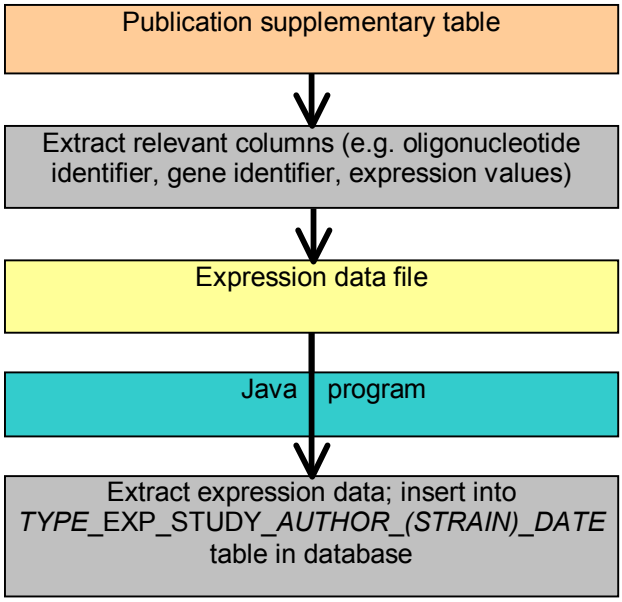


Figure 3.11. Flowchart showing the process for reading an expression dataset into the database. All expression datasets were obtained from published studies as supplementary tables.

The expression datasets are available in the form of downloadable text files as supplementary material to their respective publications. Figure 3.11 shows the

process by which the expression data was extracted and inserted into the MaGnET database.

There are two types of expression data: absolute values and ratios. As described in Chapter 1, the ratio data comes from microarrays that measure the relative intensities of spots on an experimental chip with a reference chip. These arrays are used in the study by Llinas et al. (2006) and use long oligo-nucleotides (70mer), generally one per gene, however, 990 out of 4,488 genes are represented by more than one oligo-nucleotide (Bozdech et al. 2003; Llinas et al. 2006). In this case, the expression data for multiple oligo-nucleotides cannot be averaged because it could distort the gene's true expression profile if one of the oligo-nucleotides was defective. For this type of dataset, the expression values were indexed by both the unique oligo-nucleotide identifier and the gene identifier to which it belongs.

Absolute value mRNA expression data, such as that produced in the study by Le Roch et al. (2003), is close to an approximation of the actual copy number of mRNA strands. Le Roch et al. used short (25mer) oligo-nucleotides spaced approximately every 150 base pairs on both strands of the genome and representing 5159 genes in total. They averaged the expression level of all the oligo-nucleotides for each gene. This should be less deleterious for this type of data because there were several (and for longer genes, many) short oligo-nucleotides per gene, so if one was defective it would be unlikely to distort the expression profile of the gene.

The absolute value protein expression data included in MaGnET was generated by mass spectrometry. The values represent approximate copy numbers of proteins detected. Two separate proteomic datasets were included; one comprises 2904 proteins present in at least one of seven life cycle stages (Florens et al. 2002; Le

Roch et al. 2004), the other comprises 1289 proteins in mixed asexual blood stages, gametocytes and gametes (Lasonder et al. 2002).

4. VISUALISATION PROGRAM

Overview

As described in Chapter 2, the software tool presented in this thesis aims to provide malaria biologists with innovative visualisation of various functional genomic data. Chapter 2 also introduced the main features of the tool (Figure 2.1). The tool requires a database for local data storage (described in Section 3.1) and a program for visualisation (described in this chapter).

The MaGnET visualisation program is the user interface to the datasets in the MaGnET database. The datasets include a variety of publicly available experimental genomic, transcriptomic and proteomic data with some predicted annotation (described in Section 3.2). The visualisation program encodes a set of linked viewers providing useful techniques for visualising the information. These combine to create a novel graphics-based interface that encourages users to explore genes of interest and follow-up on hunches by carrying their selections across datasets.

This chapter will describe in detail implementation of the visualisation program and will highlight important features of the user interface with the help of selected screenshots. This chapter also provides details on software availability. Lastly, the MaGnET system will be discussed and compared in relation to other tools in the field and the potential for future development will be examined.

4.1 Implementation

MaGnET was developed on a Toshiba Satellite Pro U200 laptop computer with a 1.83GHz Intel processor, 1 GB of RAM and running Microsoft Windows XP

Professional Service Pack 2 with JRE version 1.6. The MaGnET visualisation program has been written in the Java programming language, a widely-used, multi-platform language described in more detail below. One advantage of using Java was the ease of integrating third-party libraries and tools to extend the functionality of MaGnET, which, as described below, provided useful enhancements for protein structure and expression data visualisation.

4.1.1 The Java programming language

The MaGnET visualisation program is the graphical user interface to the database and, therefore, the programming language selected needed to have excellent graphics development capability. This language also needed to be cross-platform, powerful, easy to obtain and install, extensively documented, relatively simple to learn, supported by and supportive of major technologies, including the chosen RDBMS, free to use, safe and reliable.

Sun Microsystems' Java programming language (Sun Microsystems, Santa Clara, CA, USA; http://www.java.com/) was chosen because it meets the above criteria and more as one of the most important and fastest growing technologies of the current age. Java allows programmers to develop applications on one platform and run them on others, create programs to run within web browsers and to program highly customisable applications. Java is also a more secure language than other object-oriented languages, such as C++, although it is somewhat lacking in speed and efficiency.

An extensive set of libraries and resources are available to facilitate the creation of complex and powerful applications, with accompanying tutorials to lessen

the learning curve. The Java API (Application Programming Interface) includes the Abstract Window Toolkit (AWT) and the Swing toolkit for creating sophisticated graphical user interfaces (GUIs). AWT provides so called heavyweight components, such as window layout managers, which use the underlying operating system's native code to perform their functionality. Swing is built on top of AWT and offers lightweight components for GUI development, such as improved text boxes and buttons, which are purely written in Java and are truly platform independent, as well as high-level components, such as tabbed panes. The MaGnET Java program makes use of both AWT and Swing to implement the user interface.

Java is free of charge and simple to install for all users, and they can quickly check if they have it installed on their computer by visiting the Java website. Once the JRE has been installed, it will automatically check for and download updates to ensure that the user has the most recent version.

4.1.1.1 Java program and database communication

MaGnET makes use of the Java Database Connectivity (JDBC) API provided by Sun Microsystems as part of the Java distribution enabling database-independent connectivity. MySQL provides JDBC support through a driver downloadable from `http://dev.mysql.com/downloads/connector/j/5.1.html`.

4.1.2 Third-party software

Third-party software and libraries were incorporated in MaGnET to provide visualisation functionality for experimental and predicted protein structures and gene and protein expression profiles, respectively. The Jmol project (Jmol; http://www.jmol.org/) is developing a powerful, open-source Java tool for 3D

visualisation of molecular structures. By packaging Jmol classes within MaGnET, it allows Jmol to be used as an "add-on" to the main program. Users can open the Jmol program to view protein structures at the click of a button, a functionality currently not offered by any other *Plasmodium* resource in the field.

The JFreeChart library (JFreeChart; http://www.jfree.org/jfreechart/) is a free Java library for drawing custom graphs for inclusion within a Java application. The JFreeChart library was utilised to produce gene and protein expression profile graphs as line charts. Graphs have been customised to allow the display of different data types; for example, modified graphs are able to display data with multi-probe mRNA expression profiles as opposed to mRNA expression data representing probe sets.

Incorporated third-party software and their implementation are described below in more detail.

4.1.2.1 Protein structure visualisation with Jmol

Many programs exist for visualising 3D protein structures; the most widely used free programs are probably RasMol (Bernstein 2000) and its derivative browser plug-in, Chime (MDL, San Ramon, CA, USA; http://www.mdl.com/) which have now been superseded by Protein Explorer (Martz 2002). The major disadvantage of RasMol and Chime is that the user must learn a complex scripting language in order to use them effectively. Since 2002 the Jmol project (Jmol; http://www.jmol.org/) has been actively building a viable replacement for Chime. Jmol, written in Java, provides a stand-alone application and a browser applet version. Jmol is free, open-source software available under the GNU Lesser General Public License (LGPL). It has been developed for high-performance 3D rendering with no platform, hardware or web browser requirements. At the time of writing, Jmol incorporates most of

Chime's features, as well as several new ones, and is a powerful and popular viewer for molecular structures. While Jmol recognises many of the RasMol/Chime commands, which allows its behaviour to be controlled through the use of scripts, the application has an extensive menu functionality which allows non-expert users to easily interact with the displayed molecule (Jmol; http://www.jmol.org/).

The most recent stable release of the Jmol source code and executable files were downloaded from `http://sourceforge.net/project/showfiles.php?group_id=23629`. The compiled Jmol class files are included in the MaGnET binary package (JAR file) that is accessed by users operating MaGnET online. It is necessary to remove the line of code in Jmol's main class that causes the Java Virtual Machine (JVM) to close when the Jmol program is exited (this command causes the MaGnET program to also close). When users download a local version of the MaGnET software they will not receive the Jmol class files, instead they will be able to link to a downloaded version of the Jmol binary package (which allows them to maintain an up-to-date version of Jmol on their computer).

4.1.2.2 Time-series expression profile visualisation using the JFreeChart library

The JFreeChart library offers an extensive API for creating a wide variety of charts and is easy to implement and extend (JFreeChart; http://www.jfree.org/jfreechart/). Chart design is flexible so the output can easily be customised. The software is distributed under the GNU LGPL. Version 1.0.5 of the library was downloaded from `http://sourceforge.net/project/showfiles.php?group_id=154`

94. JFreeChart was used to develop the expression profile graphs available within the MaGnET Expression Data Viewer. The MaGnET charts were created using the *createLineChart* method of the ChartFactory class. Expression data were added to charts as instances of the CategoryDataset class.

4.2 Java program

The Java source code and program binaries are included on the accompanying CD.

Java is an object-oriented programming language, which is a style of programming that represents objects and their interactions by dividing the code into discrete "classes". Each class has a set of attributes and methods defining its properties and behaviour. Figure 4.1 is a class diagram for the MaGnET Java program. A class diagram depicts the organisation of the program classes and their relationships to each other. The class diagram also includes some of the major attributes of each class that define their functionality, such as windows and gene lists.

The program includes classes coding for the MaGnET data viewers: the Genome Viewer (Genome class), the Chromosome Viewer (Chromosome class), the Data Analysis Viewer (Analysis class), the Protein-Protein Interaction Viewer (PPIGraph class) and the Expression Data Viewer (Transcriptome class). Other classes provide additional functionality, such as a Gene class (to represent a gene and display its fact sheet), an ExpressionDataset class (to hold expression data), and various TimeSeriesChart classes (to display expression profile graphs). Other supporting classes represent objects used by the visualisation program, such as the ProteinStructureModel class (to represent a protein comparative structure model),

OrthologCluster (a group of homologous genes) and GeneVector (a list of user-selected genes). Classes such as DatabaseConnector, OpenJmol and BrowserLaunch take care of specific tasks as requested by the program, such as connecting to the MaGnET database, launching the Jmol viewer and loading a protein structure, and opening a web browser with a link-out to a gene in another resource, respectively. Lastly, the program includes the MAGNETMainFrame class, which contains the program's *main* method, maintains overall control of the program and holds certain globally-used attributes, such as expression dataset objects and gene lists.

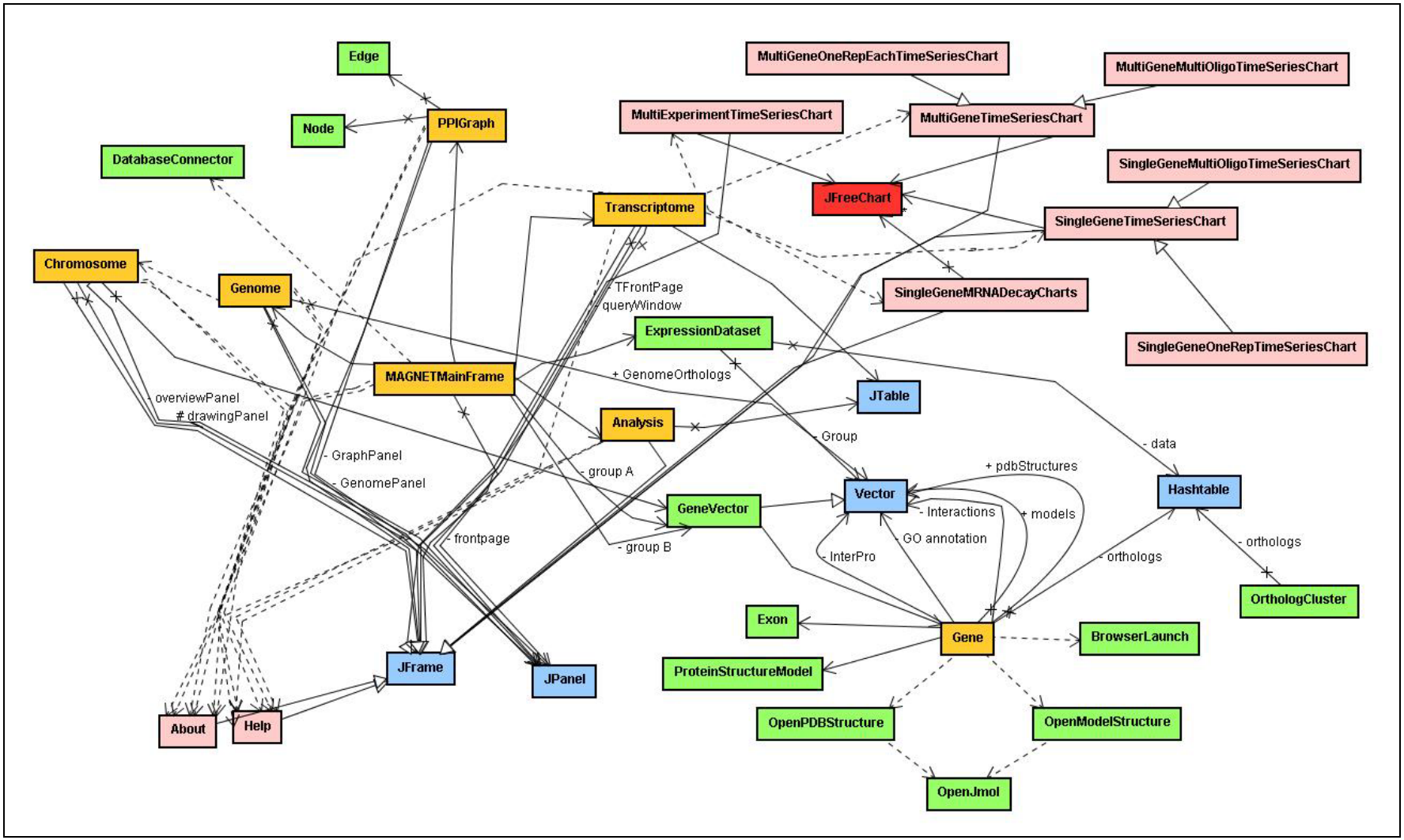

Edge
Node
PPIGraph
DatabaseConnector
MultiGeneOneRepEachTimeSeriesChart
MultiGeneMultiOligoTimeSeriesChart
MultiExperimentTimeSeriesChart
MultiGeneTimeSeriesChart
SingleGeneMultiOligoTimeSeriesChart
JFreeChart
Transcriptome
SingleGeneTimeSeriesChart
Chromosome
Genome
TFrontPage
queryWindow
SingleGeneMRNADecayCharts
SingleGeneOneRepTimeSeriesChart
ExpressionDataset
MAGNETMainFrame
+ GenomeOrthologs
JTable
- overviewPanel
drawingPanel
Analysis
- Group
- data
- GraphPanel
- GenomePanel
- group A
+ pdbStructures
Hashtable
Vector
GeneVector
- Interactions
+ models
- frontpage
- group B
- GO annotation
- InterPro
- orthologs
- orthologs
OrthologCluster
Exon
Gene
BrowserLaunch
JFrame
JPanel
ProteinStructureModel
About
Help
OpenPDBStructure
OpenModelStructure
OpenJmol

Figure 4.1. MaGnET visualisation program Unified Modelling Language (UML) class diagram. Orange boxes are the classes coding for the main MaGnET Viewers, such as the Genome Viewer. Pink boxes are classes that code for other windows within MaGnET, such as expression profile graphs. The JFreeChart library class on which all the expression profile graphs are based is coloured red. Green boxes represent visualisation program supporting classes, such as GeneVector, which holds lists of Gene objects. Blue boxes represent the major Java Swing classes that form the backbone of the MaGnET user interface (such as windows and tables). Solid lines with closed arrow heads represent generalisation – inheritance links between classes (the arrow points to the superclass). Solid lines with open arrow heads represent association – relationships between instances of two classes where one must know about the other in order to function. Dashed lines with open arrow heads represent dependency – one class depends on another and changes in one could force changes in the other. Class diagram created with the JUDE UML modelling tool, community edition (Change Vision, Inc., Tokyo, Japan; http://jude.change-vision.com/).

4.3 User interface

The MaGnET user interface includes four linked data viewers: the Genome, Protein-Protein Interaction, Expression Data and Data Analysis Viewers. A Chromosome Viewer is accessible via the Genome Viewer allowing in depth exploration of gene organisation. Pop-up gene fact sheets listing detailed information about a particular gene are available at any point. The following sections describe the program's functionality. Detailed user manuals and helpful hints can be accessed via a 'Help' menu on every page of the user interface and a tutorial is provided on the MaGnET website.

4.4.1 The MaGnET front page and MAGNETMainFrame class

The MAGNETMainFrame class contains the program's *main* method, which is called when the program starts. The *main* method calls the DatabaseConnector *connect* method, which attempts to establish a connection to the MaGnET database. If the connection is successful, the MaGnET front page is displayed.

The front page displays links to the four Data Viewers in the centre of the page. Each viewer opens in a separate window, allowing users to quickly switch between viewers and compare genes across different data types. Menu options available to the user include the option to load a file containing a list of genes (PlasmoDB standard gene identifiers are required). Users may also return to this page and save a list of their selected genes to a file, which is useful for continuing their research using other resources that accept lists of gene identifiers, or when returning to MaGnET.

Other menu options change the colours used for display of various data, such as for selected groups of genes (default colours for groups A and B are orange and blue, respectively). The ability to change the colours used by the visualisation program was a frequently requested feature at demonstrations of MaGnET. Furthermore, users can switch between the default colour scheme and a completely contrasting colour scheme by selecting a single option from the menu. The alternative colour scheme was suggested by a test-user of MaGnET with colour vision impairment. The default colour scheme uses a standard red-green scale for display of expression data, whereas the alternative colour scheme uses a yellow-purple scale. Therefore, users can select and modify the best colour scheme for their needs; a feature uniquely offered by MaGnET. A helpful future improvement would be to enable users to save their settings and automatically detect them at program start-up.

4.3.1.1 Attributes and methods of the MAGNETMainFrame class

The MAGNETMainFrame class contains the fields that need to be visible to all data viewers (known as "global" variables). Table 4.1 describes the global variables of the MAGNETMainFrame class.

Global methods defined by the MAGNETMainFrame class include methods that determine the colour a gene should be displayed to represent its expression at a particular time-point (depending on the user's preference for how the information is displayed). Three different algorithms for calculating how expression data are displayed are currently implemented in MaGnET. The default colour scheme is worked out by representing each individual gene's expression level at a given time-point by its rank within the interquartile range of the gene's expression across the

time-series experiment. Alternatively, users can choose a colour scheme based on fold change between two time-points.

Inner classes defined within the MAGNETMainFrame class are the ButtonAndMenuHandler class to deal with user requests to open particular viewers, a GroupsListener class to deal with user requests to load or save genes to a file, and a ColorChangeListener class, which displays a colour palette to assist users to change the colours of various program features.

Variable
Window height and width
Instance of Genome Viewer
Instance of Expression Data Viewer
Instance of Protein-Protein Interaction Viewer
Instance of Data Analysis Viewer
Two instances of GeneVector to hold selected groups of genes (A and B)
Colour variables for various data types
Table of orthologs/paralogs for a selected gene
Two instances of ExpressionDataset to hold a genome-wide expression dataset and expression data for sets of selected genes
User selected preferences for how expression data is displayed, such as colouring by value or by fold change

Table 4.1. Global attributes of the MAGNETMainFrame class that are accessible to all data viewers.

4.3.2 The Data Analysis Viewer and Analysis class

The Data Analysis Viewer provides tools to search the MaGnET database and to tabulate information about genes of interest. Malaria biologists using MaGnET for the first time are likely to want to search for genes they are already familiar with and to explore various data about them. Facilitating the retrieval of genes by their PlasmoDB gene identifier ensures a consistent approach to searches between MaGnET and primary genome databases, such as PlasmoDB. Providing methods to search for genes by name is clearly important too.

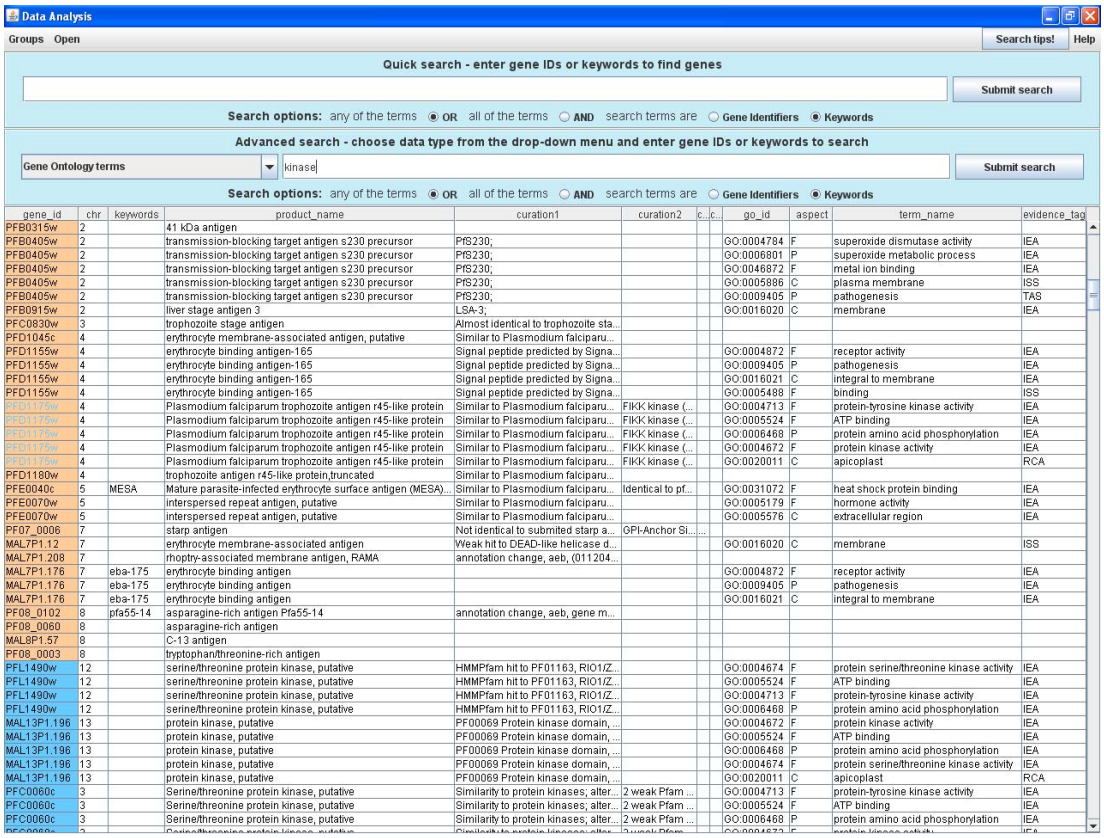
MaGnET provides a 'quick search' tool that retrieves genes by their PlasmoDB identifiers or by matching keywords to gene names. Summaries of the returned genes are displayed in a table. Selected genes can then be added to one of two groups (A and B) that act as placeholders for user-selected genes. Users can easily add or remove individual or subsets of genes to and from their selections at any time while using MaGnET. Genes selected within the Data Analysis Viewer can be carried forward to other MaGnET Data Viewers so that users can explore various aspects of gene function. Selected genes will always be identified by a particular colour (defaults are orange for group A and blue for group B).

Besides selecting genes to carry forward to other Data Viewers, users can choose to display a 'gene fact sheet' for a particular gene. Gene fact sheets are described in Section 4.3.7.

In addition to the quick search tool an 'advanced search' facility is provided for more detailed interrogation of the database. The advanced search is provided to help users to find out curated (such as that provided by the sequencing consortium) and predicted (such as Gene Ontology and InterPro) functional annotations for genes they have become interested in during the course of exploration using MaGnET. There are two options for retrieving functional annotation: retrieving the annotation for specific genes using their gene identifiers and searching within annotation of all genes for matches to particular keywords.

Further functionality provided by the Data Analysis Viewer is the display of a summary table for all genes in user-selected groups A and B (Figure 4.2). The table includes gene names, curated annotation and GO terms. Tabular display of data can help a user to notice trends in the annotation; for example the repeated occurrence of

a particular GO term can provide another branch of evidence that there may be a linked functionality between the genes that is worth exploring further.



gene_id	chr	keywords	product_name	curation1	curation2	c...	c...	go_id	aspect	term_name	evidence_tag
PFB0315w	2		41 kDa antigen								
PFB0405w	2		transmission-blocking target antigen s230 precursor	Pfs230;				GO:0004784	F	superoxide dismutase activity	IEA
PFB0405w	2		transmission-blocking target antigen s230 precursor	Pfs230;				GO:0006801	P	superoxide metabolic process	IEA
PFB0405w	2		transmission-blocking target antigen s230 precursor	Pfs230;				GO:0046872	F	metal ion binding	IEA
PFB0405w	2		transmission-blocking target antigen s230 precursor	Pfs230;				GO:0005886	C	plasma membrane	ISS
PFB0405w	2		transmission-blocking target antigen s230 precursor	Pfs230;				GO:0009405	P	pathogenesis	TAS
PFB0915w	2		liver stage antigen 3	LSA-3;				GO:0016020	C	membrane	IEA
PFC0830w	3		trophozoite stage antigen	Almost identical to trophozoite sta...							
PFD1045c	4		erythrocyte membrane-associated antigen, putative	Similar to Plasmodium falciparu...							
PFD1155w	4		erythrocyte binding antigen-165	Signal peptide predicted by Signa...				GO:0004872	F	receptor activity	IEA
PFD1155w	4		erythrocyte binding antigen-165	Signal peptide predicted by Signa...				GO:0009405	P	pathogenesis	IEA
PFD1155w	4		erythrocyte binding antigen-165	Signal peptide predicted by Signa...				GO:0016021	C	integral to membrane	IEA
PFD1155w	4		erythrocyte binding antigen-165	Signal peptide predicted by Signa...				GO:0005488	F	binding	ISS
PFD1175w	4		Plasmodium falciparum trophozoite antigen r45-like protein	Similar to Plasmodium falciparu...	FIKK kinase (...			GO:0004713	F	protein-tyrosine kinase activity	IEA
PFD1175w	4		Plasmodium falciparum trophozoite antigen r45-like protein	Similar to Plasmodium falciparu...	FIKK kinase (...			GO:0005524	F	ATP binding	IEA
PFD1175w	4		Plasmodium falciparum trophozoite antigen r45-like protein	Similar to Plasmodium falciparu...	FIKK kinase (...			GO:0006468	P	protein amino acid phosphorylation	IEA
PFD1175w	4		Plasmodium falciparum trophozoite antigen r45-like protein	Similar to Plasmodium falciparu...	FIKK kinase (...			GO:0004672	F	protein kinase activity	IEA
PFD1175w	4		Plasmodium falciparum trophozoite antigen r45-like protein	Similar to Plasmodium falciparu...	FIKK kinase (...			GO:0020011	C	apicoplast	RCA
PFD1180w	4		trophozoite antigen r45-like protein,truncated	Similar to Plasmodium falciparu...							
PFE0040c	5	MESA	Mature parasite-infected erythrocyte surface antigen (MESA)...	Similar to Plasmodium falciparu...	Identical to pf...			GO:0031072	F	heat shock protein binding	IEA
PFE0070w	5		interspersed repeat antigen, putative	Similar to Plasmodium falciparu...				GO:0005179	F	hormone activity	IEA
PFE0070w	5		interspersed repeat antigen, putative	Similar to Plasmodium falciparu...				GO:0005576	C	extracellular region	IEA
PF07_0006	7		starp antigen	Not identical to submitted starp a...	GPI-Anchor Si...						
MAL7P1.12	7		erythrocyte membrane-associated antigen	Weak hit to DEAD-like helicase d...				GO:0016020	C	membrane	ISS
MAL7P1.208	7		rhoptry-associated membrane antigen, RAMA	annotation change, aeb, (011204...							
MAL7P1.176	7	eba-175	erythrocyte binding antigen					GO:0004872	F	receptor activity	IEA
MAL7P1.176	7	eba-175	erythrocyte binding antigen					GO:0009405	P	pathogenesis	IEA
MAL7P1.176	7	eba-175	erythrocyte binding antigen					GO:0016021	C	integral to membrane	IEA
PF08_0102	8	pfa55-14	asparagine-rich antigen Pfa55-14	annotation change, aeb, gene m...							
PF08_0060	8		asparagine-rich antigen								
MAL8P1.57	8		C-13 antigen								
PF08_0003	8		tryptophan/threonine-rich antigen								
PFL1490w	12		serine/threonine protein kinase, putative	HMMPfam hit to PF01163, RIO1/Z...				GO:0004674	F	protein serine/threonine kinase activity	IEA
PFL1490w	12		serine/threonine protein kinase, putative	HMMPfam hit to PF01163, RIO1/Z...				GO:0005524	F	ATP binding	IEA
PFL1490w	12		serine/threonine protein kinase, putative	HMMPfam hit to PF01163, RIO1/Z...				GO:0004713	F	protein-tyrosine kinase activity	IEA
PFL1490w	12		serine/threonine protein kinase, putative	HMMPfam hit to PF01163, RIO1/Z...				GO:0006468	P	protein amino acid phosphorylation	IEA
MAL13P1.196	13		protein kinase, putative	PF00069 Protein kinase domain, ...				GO:0004672	F	protein kinase activity	IEA
MAL13P1.196	13		protein kinase, putative	PF00069 Protein kinase domain, ...				GO:0005524	F	ATP binding	IEA
MAL13P1.196	13		protein kinase, putative	PF00069 Protein kinase domain, ...				GO:0006468	P	protein amino acid phosphorylation	IEA
MAL13P1.196	13		protein kinase, putative	PF00069 Protein kinase domain, ...				GO:0004674	F	protein serine/threonine kinase activity	IEA
MAL13P1.196	13		protein kinase, putative	PF00069 Protein kinase domain, ...				GO:0020011	C	apicoplast	RCA
PFC0060c	3		Serine/threonine protein kinase, putative	Similarity to protein kinases; alter...	2 weak Pfam ...			GO:0004713	F	protein-tyrosine kinase activity	IEA
PFC0060c	3		Serine/threonine protein kinase, putative	Similarity to protein kinases; alter...	2 weak Pfam ...			GO:0005524	F	ATP binding	IEA
PFC0060c	3		Serine/threonine protein kinase, putative	Similarity to protein kinases; alter...	2 weak Pfam ...			GO:0006468	P	protein amino acid phosphorylation	IEA

Figure 4.2. Screenshot of the MaGnET Data Analysis Viewer. The table is displaying a summary of information about the genes that have been added to two groups, A (orange) and B (blue). Genes appearing in both groups are represented by an orange background with blue text. The data summarised in the table include genomic location, product name, curated annotation and GO annotation. Each GO annotation is displayed on a separate row; therefore a single gene may span multiple rows.

It should be noted here that MaGnET does not provide quantitative analysis of significant GO term over-representation in a gene list (many tools exist to do this). However, qualitative analysis like that encouraged by MaGnET can be useful for finding interesting leads for further investigation relatively quickly and simply.

Other advantages will be the ability to pick up by eye on smaller or short-lived changes that are often missed by statistics-based algorithms.

4.3.2.1 Attributes and methods of the Analysis class

Attributes of the Analysis class include various buttons and text fields to capture input search options and an instance of the Java JTable class to display the returned genes in tabular format. Database searching is handled by a ButtonHandler inner class that responds to a user pressing the 'Search' button. A GroupsMenuHandler inner class deals with adding, removing and displaying genes in the selected groups.

4.3.3 Genome Viewer and Genome class

The Genome Viewer facilitates the display of genomic information, such as location of protein encoding genes, pseudogenes and RNA encoding genes. The Genome Viewer displays the 14 nuclear chromosomes as vertical bars drawn from left to right on the screen according to increasing chromosome number (as assigned by the genome sequencing consortium). It also displays the circular apicoplast chromosome and the mitochondrial chromosome. The nuclear chromosomes are drawn in proportion to one another, with the longest chromosome scaled to the screen height at the lowest zoom level. The organellar genomes are drawn at a scale 5 times greater than the other chromosomes because if they were drawn at the same scale they would be too small to be usefully rendered.

Upon opening the Genome Viewer automatically displays the locations of all genes in selected groups A and B (Figure 4.3). A useful legend is provided describing the colour scheme used to display various genomic features. Another

helpful attribute is a search bar for quickly locating genes using their standard gene identifiers.

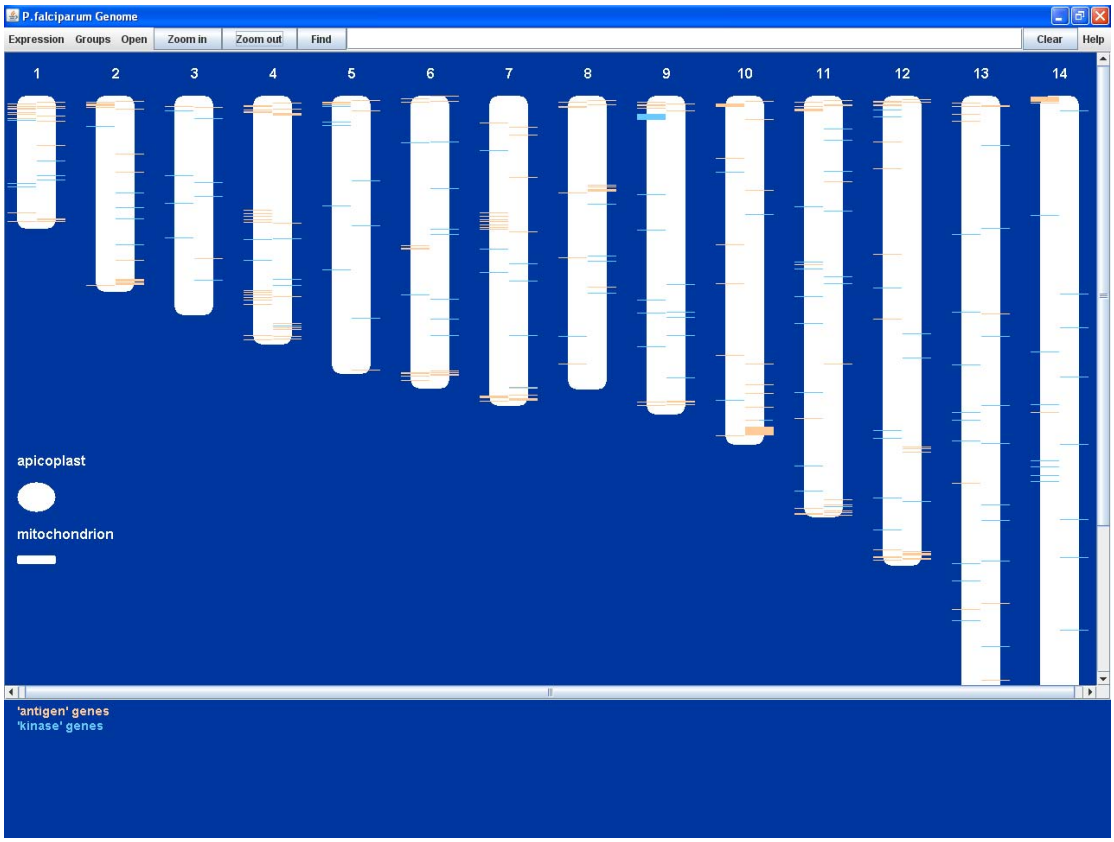


Figure 4.3. Screenshot of the MaGnET Genome Viewer. Two groups of genes selected from keyword searches are displayed as coloured lines ('antigen' in orange and 'kinase' in blue). Genes on the forward strand are drawn on the right side of the chromosome; genes on the reverse side are drawn on the left. Genes in both groups are drawn as split lines with the left half coloured orange and the right half coloured blue.

One of the advanced features that MaGnET provides is the ability to overlay expression data onto genomic location. Users can visualise an expression dataset chosen from a menu. If the user has already selected genes and stored them in groups A or B, display of expression data can be optionally limited to just the genes

in these groups (Figure 4.4). Alternatively, a genome-wide expression dataset can be displayed (Figure 4.5).

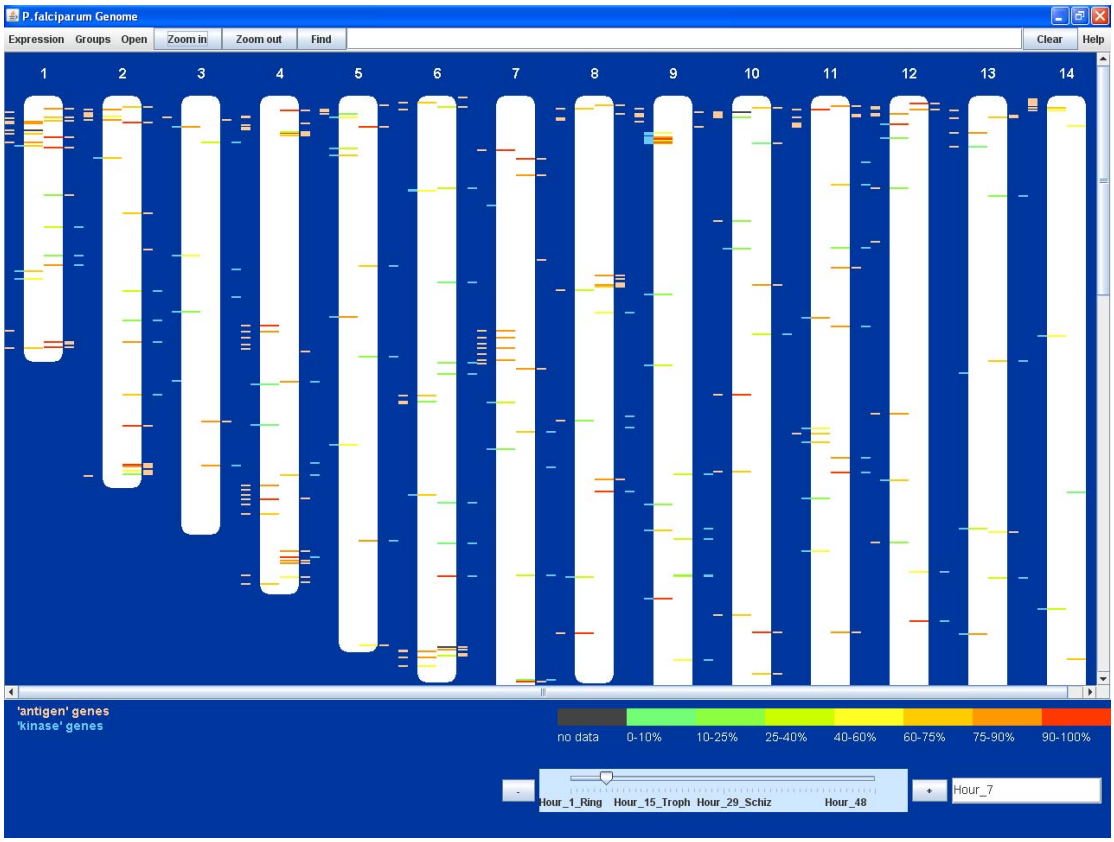


Figure 4.4. Screenshot of the Genome Viewer displaying mRNA expression data for genes in two selected groups. The genes are coloured according to rank at hour 7 within the interquartile range of each gene's expression across the time-series experiment [in this case the intraerythrocytic development cycle (IDC)]. Lines marking the location of genes in the selected groups are drawn to the sides of the chromosomes. The display has been magnified relative to Figure 4.3 by using the 'Zoom in' button. A slider in the lower panel allows users to step-through the time-series.

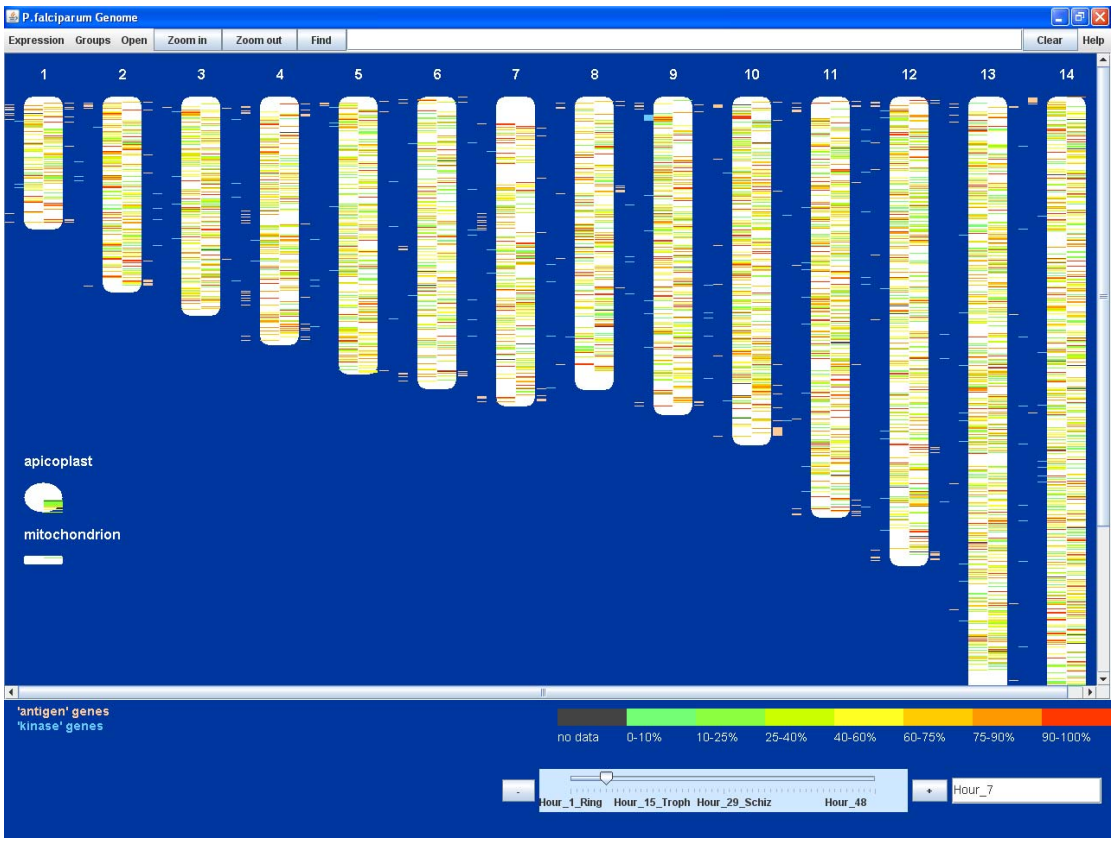


Figure 4.5. Screenshot of the Genome Viewer displaying an mRNA expression dataset mapped onto genomic location of the genes. The genes are coloured according to rank at hour 7 within the interquartile range of each gene's expression across the IDC. A slider in the lower panel allows users to step-through the time-series.

Expression data for a single experimental time-point are displayed at any one time and a useful feature is the provision of a slider to enable users to move through a time-series. This allows users to explore expression changes over time in relation to genomic location. Locally correlated patterns of expression changes can be an indication of co-regulation and likely shared function (Cohen et al. 2000).

Section 4.3.1.1 describes the options available for expression data representation. Simply, option one represents expression data at the current time-point relative to the range expression seen across the whole time-series (Figures 4.4

and 4.5). Option two represents directionality and magnitude of changes in expression between previous and current time-points (Figure 4.6). The latter is particularly useful because it simplifies the display of expression data by highlighting only genes that are significantly (greater than two fold) up- or down-regulated between time-points. This may make trends easier to notice than by looking at the complicated pattern of mRNA abundances provided by the first option.

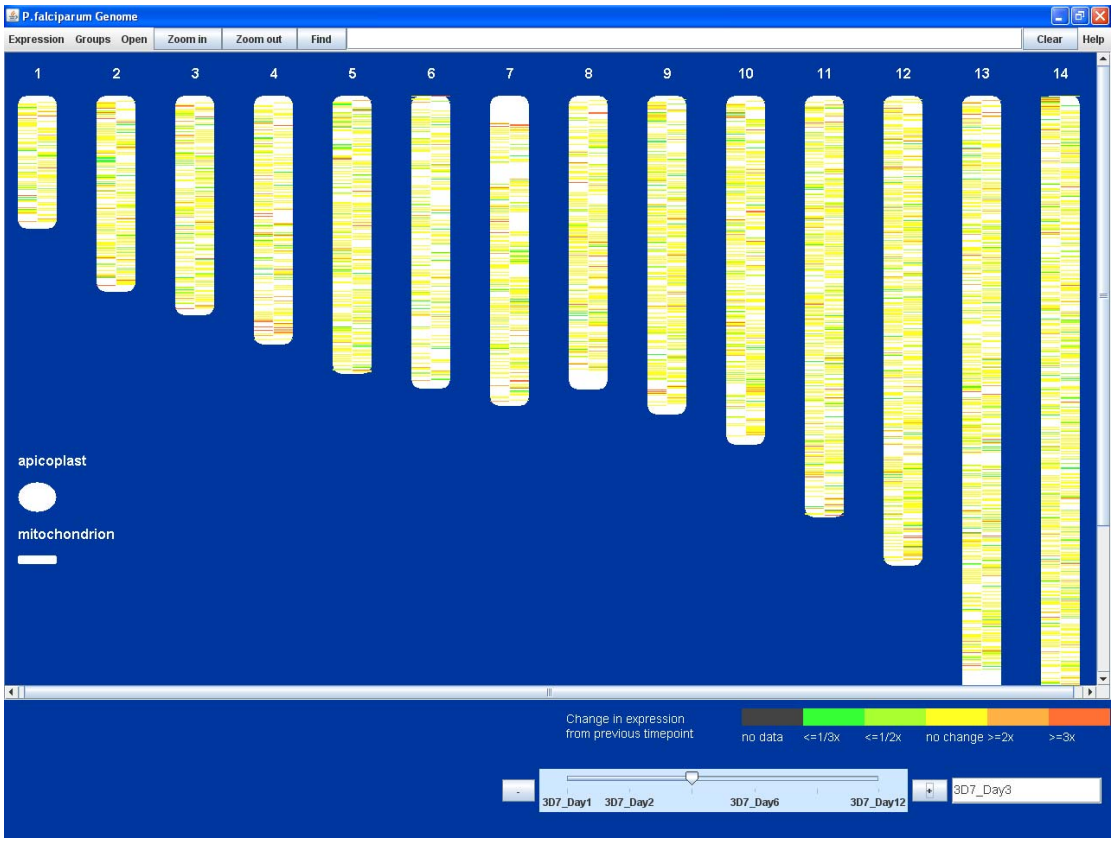


Figure 4.6. Screenshot of the Genome Viewer displaying the direction of changes in mRNA expression from the previously sampled time-point. The genes are coloured according to their change in expression from gametocyte days two to three. A significant increase in expression is coloured orange and a significant decrease is coloured green [a change greater than 2 fold is considered significant (Li et al. 2005)].

MaGnET facilitates exploration of gene families through features available within the Genome and Chromosome Viewers. Gene families are selected within the Chromosome Viewer, although their genome-wide localisation is visualised via the Genome Viewer. This would be useful, for example, to investigate expression patterns of gene families over various life cycle stages. Visualisation of gene families is described further in Section 4.3.4.2.

4.3.3.1 Attributes and methods of the Genome class

Attributes of the Genome class include the various drawing panels, buttons, text field and slider required by the Viewer, as well as the chromosome dimensions and a Vector object to store a list of paralogous genes for display of gene families. The Genome class gets access to information about genes in selected groups and expression data from globally available attributes and methods of the MaGnETMainFrame class. Inner classes include a ClickOnChrHandler that listens for a user double clicking on a chromosome in order to launch the Chromosome Viewer. A ScrollingHandler deals with repainting the viewing panel when a user moves the scroll bar. An ExpressionMenuHandler responds to user requests to load expression datasets, to switch on or off the option for display of data for selected genes versus whole datasets, and to change the way that expression data are represented. A SliderChangeListener recalculates and repaints displayed expression data when a user moves between time-points using the slider. A ButtonHandler inner class has several functions, including responding to user requests to locate genes from the search bar and to zoom in and out on the display panel. The RedrawOnFocus inner class allows the Genome Viewer to automatically detect and

display changes in selected genes or loaded expression dataset that were applied in other open viewers when focus is returned to the Genome Viewer.

4.3.4 Chromosome Viewer and Chromosome class

A Chromosome Viewer is provided to allow users to zoom in on a particular chromosome while using the Genome Viewer (Figure 4.7). The Chromosome Viewer facilitates navigation at varying levels of magnification; the achievable magnification range is between 100 base pairs per pixel and 10 base pairs per pixel.

Unlike most other genome viewers, MaGnET does not offer genome browsing at the individual base pair level. The reason for this omission was so that MaGnET could provide a simplified view of the genome in order to avoid a sensation of "data overload" likely to be felt by non-bioinformatics specialist users.

Unlike the other MaGnET Viewers, multiple Chromosome Viewer windows can be open at the same time. This facilitates the comparison of regions from two different chromosomes at the same time (or two regions of the same chromosome by repeatedly opening a particular chromosome), which would be useful for comparing gene order or gene expression patterns in regions that are linked by duplication and rearrangement, for instance.

As well as providing a zoom-able, scrollable view of the chromosome, an overview of the entire chromosome is displayed in a panel below the former. This feature allows the user to keep track of where in the chromosome the "viewing window" is currently focused. It also allows users to visualise patterns over larger regions of a chromosome, such as relationships between different multi-gene families.

Figure 4.7. Screenshot of the Chromosome Viewer. The upper panel displays a detailed view of the chromosome (the forward strand is represented by the upper bar and the reverse strand is represented by the lower bar). The centre panel shows an overview of the chromosome with the current position of the chromosome viewer represented by the grey region. The lower panel contains a legend for the colour scheme. Two groups of selected genes are displayed ('antigen' in orange and 'kinase' in blue). A gene appearing in both groups would have its upper half coloured orange and lower half coloured blue. Unselected protein-coding genes are shown in medium-grey. Unselected pseudogenes and RNA genes are shown in light grey. Introns appear as pink regions within genes. Clicking with the mouse once on a gene highlights it in the 'picking colour' (purple) and displays its product name in the top left corner of the window.

Expression data can be viewed in the Chromosome Viewer as in the Genome Viewer (see Section 4.3.3 for more details on the options available for expression data visualisation). Close-up exploration of expression patterns offers opportunities

for discovery of local patterns of co-expression, which may indicate co-regulation and linked functionality of genes. For example, Figure 4.8 shows a region on the left arm of chromosome one where several adjacent genes are maximally expressed during hour 28 of the IDC (late trophozoite stage).

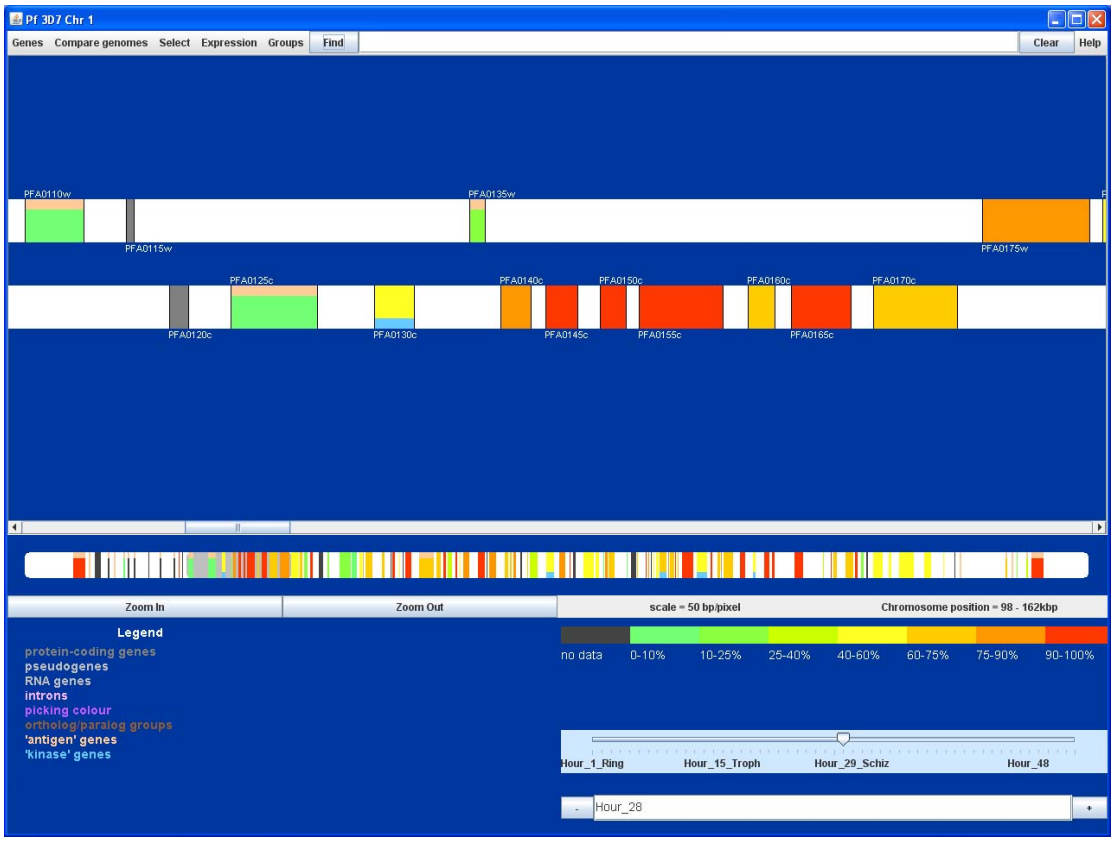


Figure 4.8. Screenshot of the Chromosome Viewer displaying an mRNA expression dataset. The genes are coloured according to rank at hour 28 within the interquartile range of each gene's expression across the IDC. Genes that appear in group A ('antigen') are marked by an orange band at the top of the gene, whereas genes appearing in group B ('kinase') are marked by a blue band at the base of the gene. A slider in the lower panel allows users to step-through the time-series.

4.3.4.1 Attributes and methods of the Chromosome class

Attributes required by the Chromosome Viewer include panels representing the scrolling chromosome viewing window, chromosome overview and legend, buttons and menus with various options including selection boxes so users can turn on or off the display of genomic features, such as introns and pseudogenes. Other display-determining attributes include the current zoom level, chromosome length (number of base pairs) and coordinates of the visible region. Several lists of genes are also required, which are: a list of all genes on the chromosome, a list of genes that are currently selected by the user (prior to being saved as a group), and a list of family members for the chromosome needed when a user is looking at gene families (see Section 4.3.4.2). A method called *GetOrthologsForChr* returns a list of family members located on the current chromosome when the user clicks on a gene whilst under gene family browsing mode (see Section 4.3.4.2).

Inner classes include SliderChangeListener which listens for the user moving the slider to move between time-points whilst visualising expression data. A ButtonHandler inner class deals with user requests via buttons, including zooming and locating genes by identifier using a search bar at the top of the Chromosome Viewer window. A RedrawOnFocus inner class allows the Chromosome Viewer to automatically detect and display changes in selected genes or loaded expression dataset that were applied in other open viewers when focus is returned to the current window. A ScrollingHandler handles repainting of the visible chromosome region when the user scrolls along the chromosome. Similarly, GenesMenuHandler deals with repainting the Chromosome Viewer if genomic features are made visible or invisible using menu options. The GroupMenuHandler inner class has responsibility

for dealing with user requests to add or remove genes from selected groups A and B. Users can easily add or remove multiple genes to and from groups by clicking on them, thereby highlighting them purple, and choosing from various menu options. The ClickOnDrawingPanelHandler retrieves the coordinates when a user clicks on the chromosome viewing panel and matches them against the list of gene coordinates to find out which gene the user clicked on and then handles the requested behaviour, such as selecting or deselecting the gene. As in the Genome Viewer, the ExpressionMenuHandler deals with requests to load expression data.

4.3.4.2 Visualising gene families

Further functionality provided by the Chromosome and Genome Viewers facilitates exploration of gene families (orthologs and paralogs). Under gene family browsing mode, when a user selects any gene in the Chromosome Viewer, all paralogous genes will automatically be highlighted in a distinct colour within the genome (Figure 4.9). A table containing number of family members present in the *P. falciparum* genome and four related genomes is displayed in the legend panel of the Chromosome Viewer. This table provides the user with an indication of how wide-spread the family is, such as whether it is unique to *P. falciparum* or has undergone expansion in this genome. A list of family members in the five genomes and links to their gene pages in PlasmoDB are provided via the gene fact sheet.

In order to facilitate users to carry forward complete groups of paralogs to other viewers, in gene family browsing mode users can add a family to one of selected groups A and B at the click of a button, eliminating the requirement to select family members individually.

Figure 4.9. Screenshots of the Genome (a) and Chromosome (b) Viewers displaying the ortholog/paralog group for gene PFA0625w. Paralogous genes are highlighted on the genome in green. A table in the lower panel lists the number of family members that are found in this and several other species of *Plasmodium*.

4.3.5 Protein-Protein Interaction Viewer and PPIGraph class

Protein-protein interactions are represented by lines (interactions) connecting nodes (proteins) in the Protein-Protein Interaction Viewer. The PPIGraph class displays interaction networks using an algorithm adapted by Richard Orton for the Yeast Exploration Tool Integrator (YETI) (Orton et al. 2004; Orton 2006) from work by Ralf Mrowka (Mrowka 2001), which was originally based on the Graph class provided with the Java Development Kit (JDK) from Sun Microsystems (Sun Microsystems, Santa Clara, CA, USA; http://www.java.com/). A relaxation algorithm is used, which constricts interacting proteins to a pre-defined distance (user adjustable) but maximises the distance between non-interacting proteins in the available two-dimensional space (Mrowka 2001).

When a user opens the Protein-protein Interaction Viewer, the interactions for any proteins stored in user selected groups A and B will automatically be fetched and displayed in the viewer. Users may also easily search for interactions for particular proteins of interest by entering one or more standard gene identifiers in a search bar at the bottom of the window. By default, all proteins that directly interact with the given protein are displayed ('primary interactions'). Users can choose to extend the network outwards to draw in other proteins that interact with this first layer of interacting proteins (which will be 'secondary interacting' proteins in relation to the original protein at the centre of the network) (Figure 4.10). Furthermore, the size of network that can usefully be displayed is only limited by the computational power and screen size of the user's computer. Therefore, protein-protein interaction network visualisation using MaGnET is theoretically only limited by the size of the

currently sampled interaction space (only one yeast two-hybrid dataset is available and this type of data often contains a large number of false negatives).

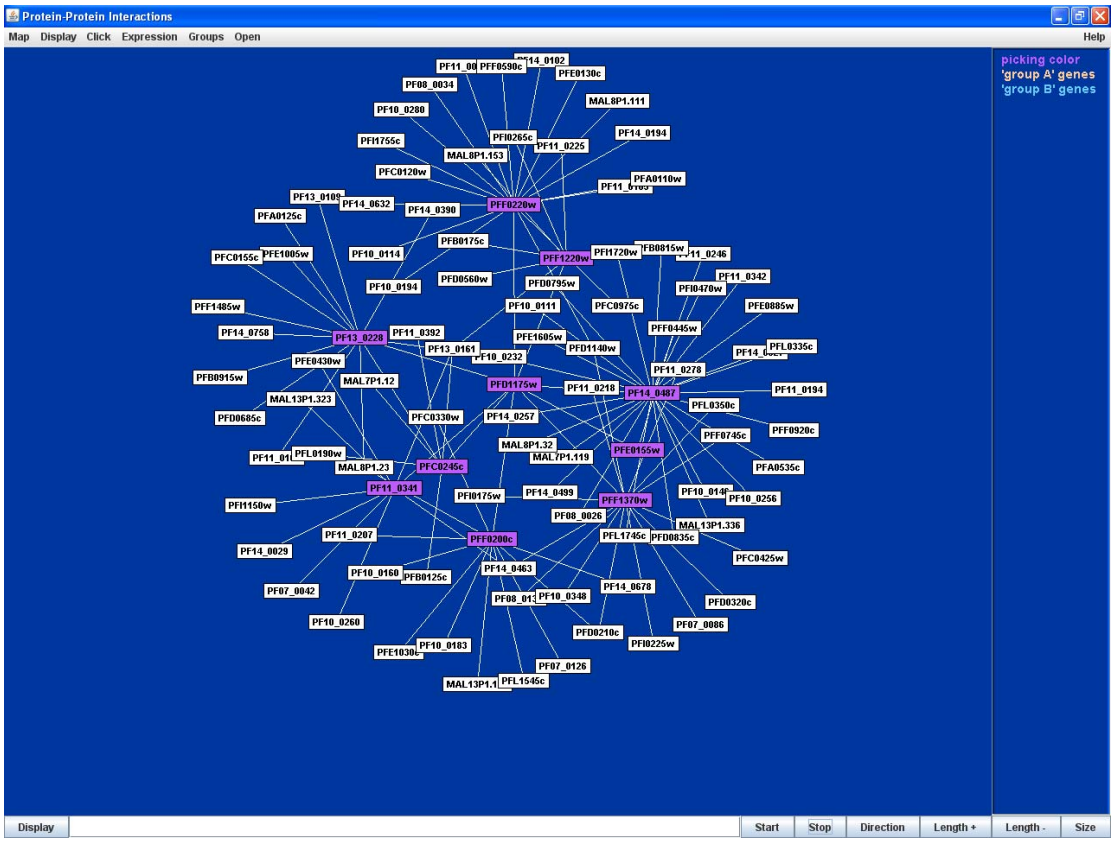


Figure 4.10. Screenshot of the primary and secondary interaction network of the R45 antigen (PFD1175w). R45 and its primary interacting proteins are highlighted in purple. All secondary interacting proteins are in white. An interaction is represented as a white line connecting two protein labels.

Advanced options for manipulating the display of interactions are provided by a menu system and a set of buttons. Useful functionality includes highlighting clusters of interacting proteins by clicking on a particular protein (Figure 4.10); the cluster can then be added to either of selected groups A or B. Another option reduces the complexity of the display by minimising the size of some nodes, while

retaining others at full size, in order to facilitate exploration of sub-regions of the network (Figure 4.11).

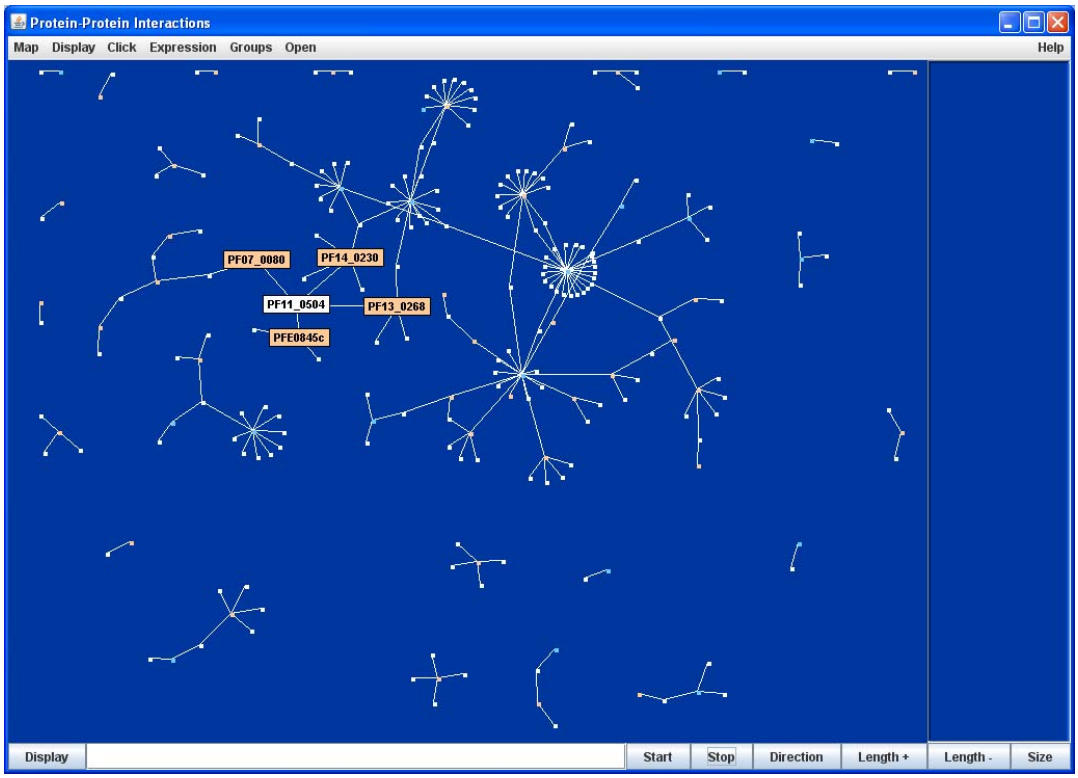


Figure 4.11. Screenshot of a protein interaction network where the majority of protein labels have been minimised but one region displays expanded protein labels. This is useful for minimising screen crowding when there are many proteins in the network.

Further advanced functionality for overlay of expression data onto protein-protein interaction network is provided (Figure 4.12). Options available to the user for display of expression data are the same as described previously for the Genome and Chromosome Viewers. As in other viewers, a useful legend is provided in the window describing the colour scheme used and housing a slider to allow movement through the time-series.

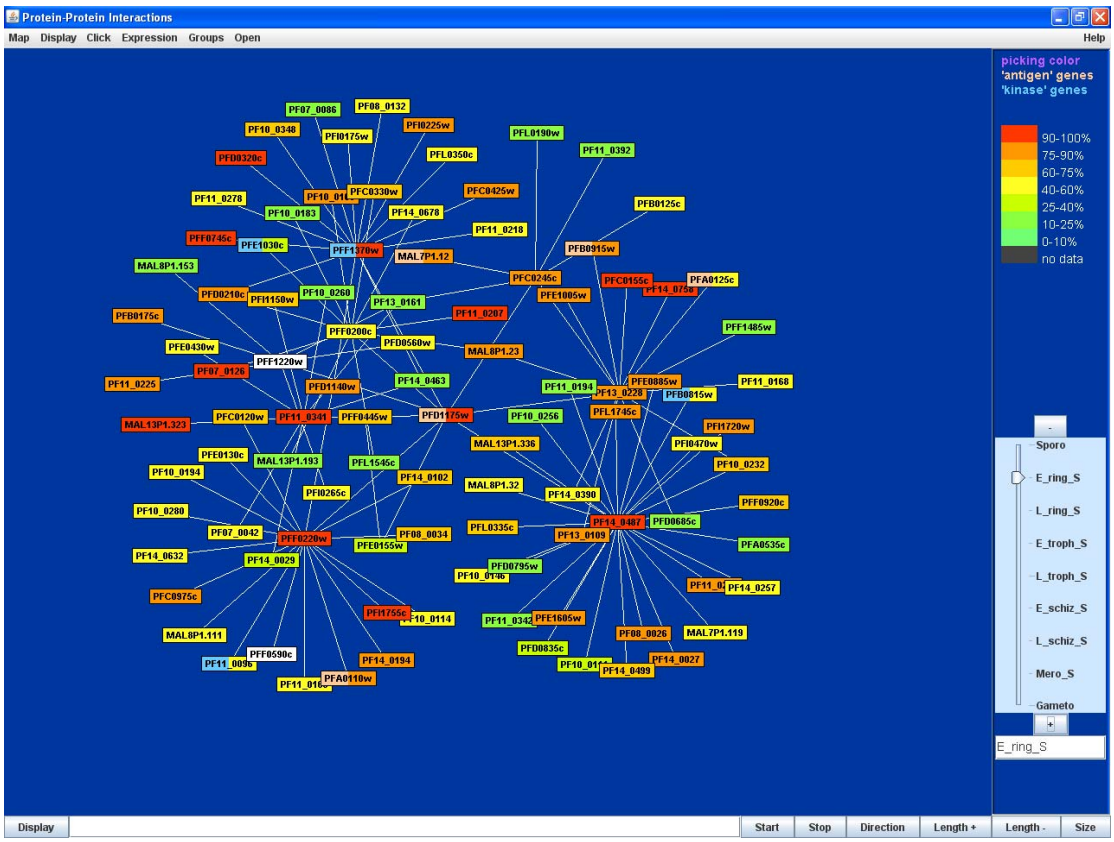


Figure 4.12. Screenshot of the R45 antigen's protein interaction network, with protein labels coloured according to their mRNA expression level at the early ring stage of the parasite's life cycle. Labels of proteins appearing in pre-selected groups are coloured half according to their group ('antigen', orange, or 'kinase', blue) and half according to their expression level. Proteins that are not represented in the expression dataset are drawn in white.

Display of expression data mapped onto protein interaction network affords various opportunities. For example, expression patterns can add weight to the likelihood of individual interactions being true positives (if the proteins are expressed during the same stage of the lifecycle), or conversely, can indicate that the interaction could be a false positive (if proteins are never expressed at the same time). Also, exploration of co-expression patterns over networks can suggest the processes that

hypothetical proteins may be involved in or indicate novel additional functions for known proteins.

While browsing the interaction network it is important to be able to access known functional information about the proteins. To facilitate this, MaGnET provides quick access to certain functional data at a click of the mouse. Users can display the protein name and GO cellular component terms associated with a protein in the text field at the bottom of the window. Also, gene fact sheets can be easily accessed by clicking on individual proteins.

Overall, the MaGnET Protein-Protein Interaction Viewer offers functionality not available in other *Plasmodium* resources. Other tools in the field, such as PlasmoDB (Bahl et al. 2003), merely provide lists of interacting partners for individual genes. None of them offer advanced tools for network browsing beyond the level of primary interactions. This disadvantages users who do not have the opportunity to explore interactions on a deeper level; for example, to discover proteins that have multiple interactions in common, a good indicator of related function (Koegl and Uetz 2007).

4.3.5.1 Attributes and methods of the PPIGraph class

The PPIGraph class contains an inner class, GraphPanel, which extends the JPanel class and implements the algorithm that displays the protein interaction graph. The GraphPanel class contains methods to add nodes (proteins) and edges (interactions), find nodes that the user has selected, implement the relaxation algorithm, and finally, to paint the resulting network in the display window.

Further inner classes represent nodes (Node class) and edges (Edge class). The RedrawOnFocus class is responsible for redrawing the window when focus is

returned (usefully re-colours proteins that have been selected or de-selected in other viewers and checks for changes to the display of loaded expression data). A ButtonHandler class deals with all commands mediated by buttons, such as freezing movement of the network map or incrementing edge-length. An ItemHandler class deals with requests via the menu options, usually switching on or off different modes of usage, such as what happens when a user clicks on a protein; for example, extending the network by bringing in other interaction partners or highlighting clusters of interacting proteins. It also deals with requests to add or remove genes from selected groups A and B. As described previously, the ExpressionDataMenuHandler deals with requests to load or alter the display of expression data and the SliderChangeListener updates expression data display when a user moves between time-points using the slider.

The PPIGraph constructor defines an instance of the Java AWT MouseAdaptor class for responding to user clicks on the drawing panel. The outcome of the action depends on defined settings, which are selectable via a set of menu options and buttons. An instance of the MouseMotionAdapter class is also defined that responds to a user dragging nodes across the drawing panel to rearrange them.

Attributes of the PPIGraph class include the graph drawing panel (GraphPanel), a legend panel (JPanel), an expression data slider (JSlider), a text field (JTextField) and various buttons and menu options. Other attributes include various options determining aspects of the display, such as length of edges and colours of nodes and edges, and arrays and vectors to hold lists of all edges and nodes, user-selected nodes and nodes whose labels have been minimised or maximised. The class

also includes a random number generator that is involved in placing the nodes onto the display panel before they are sorted by the relaxation algorithm.

Methods defined by the PPIGraph class include a *showGroups* method, which checks for interactions for proteins in selected groups A and B upon loading of the Protein-Protein Interaction Viewer and when focus is regained. The *newMap* method prepares a new interaction network given the results of a search for interactions.

4.3.6 Expression Data Viewer and Transcriptome class

The Expression Data Viewer is coded for by the Transcriptome class. 'Transcriptome' is really a misnomer because this class can deal with any kind of expression data. The class was developed before it was decided to include protein expression data in addition to mRNA expression data.

The Expression Data Viewer has two functions: it provides users with an advanced, easy-to-use search facility for querying expression data and provides visualisation of time-series profiles for individual genes or small groups of genes. The next two sections describe its functionality in detail.

4.3.6.1 Time-series graphs

The Expression Data Viewer complements the expression data visualisation capabilities of other MaGnET viewers by providing graphical displays of time-series profiles. Time-series profile graphs are displayed by typing gene identifiers into a search bar on the Expression Data Viewer window. Graphs can display data for individual genes (Figure 4.13), or for small groups of genes (Figure 4.14). A helpful alternative to typing in lists of identifiers is the option to quickly access a graph

displaying profiles for all genes in selected groups A and/or B by typing the letter(s) into the search bar. The size of group that can be usefully displayed depends on the size of the computer screen and the type of data being displayed (where the data is from glass-slide arrays, profiles are displayed for each oligo, whereas data from Affymetrix arrays and protein expression data are displayed on a per-gene basis – see Sections 1.5.1 and 1.5.2 and Table 3.1 for more information).

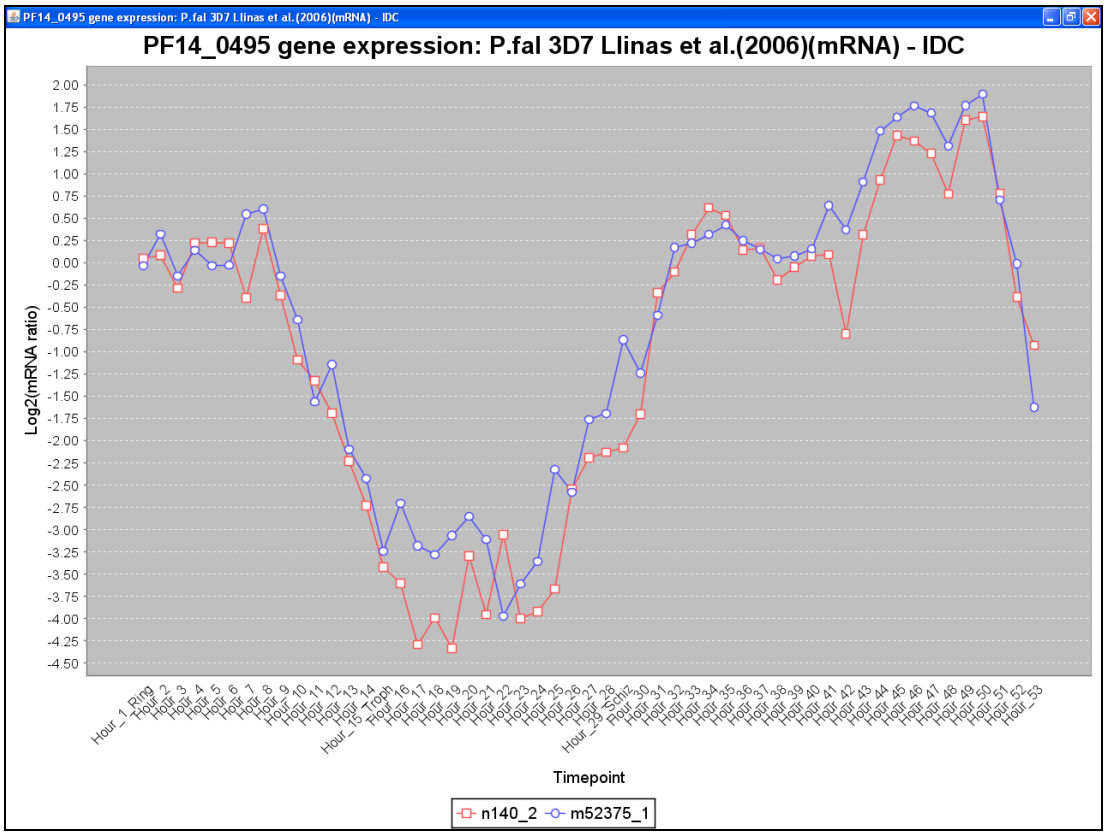


Figure 4.13. Screenshot of the time-series profile graph for the *P. falciparum* 3D7 gene PF14_0495 during the IDC (data have been $\log_2$ transformed). This gene is represented by two oligonucleotides in the array used by Llinas et al (2006). The recorded expression of each oligonucleotide is represented as a differently coloured line on the graph.

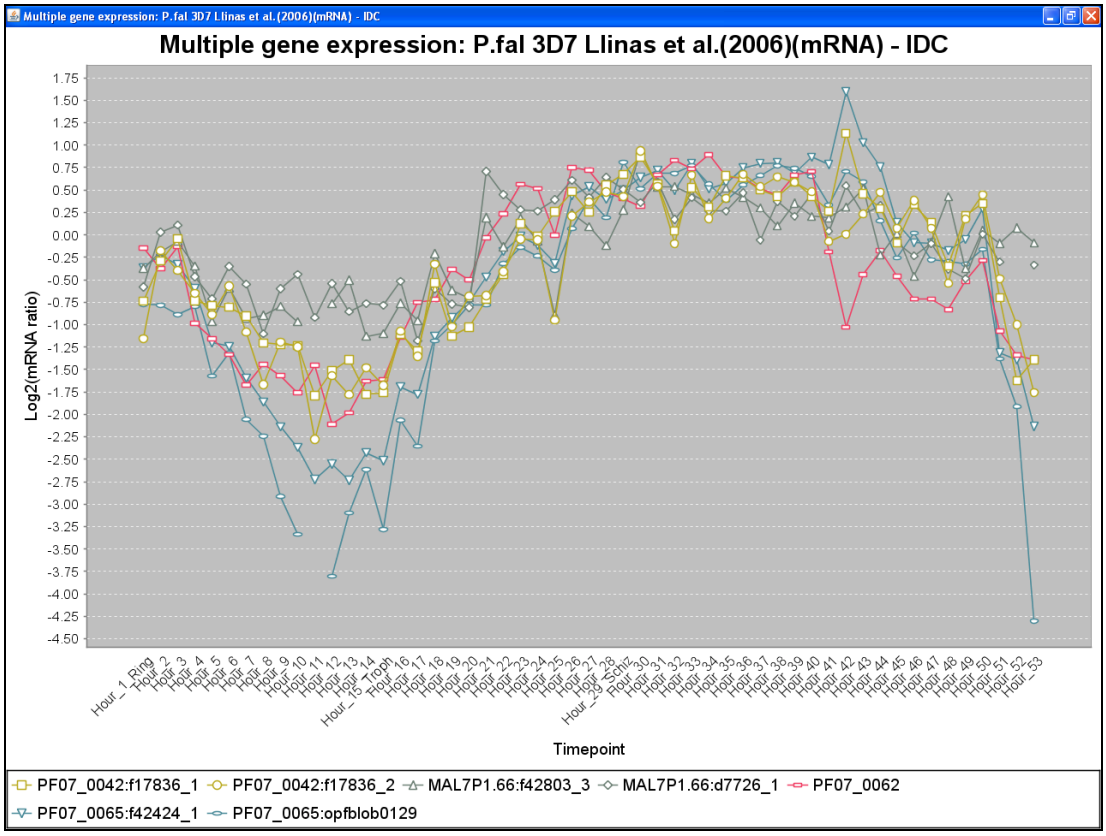


Figure 4.14. Screenshot of the time-series profile graph for four genes expressed during the IDC. Each gene is represented by a different colour and oligonucleotides are distinguished by different symbols.

Further options for display of time-series data include displaying a gene's expression profile across multiple datasets on a single graph and displaying mRNA decay rates alongside mRNA expression data from the same life cycle stage. The former provides a simple way to compare expression between datasets; however the current implementation is limited and provides opportunity for improvement in future versions. The latter option is useful to investigate how mRNA decay rates relate to observed expression; a high decay rate but increasing expression indicate that a gene is being actively transcribed at the current stage (Figure 4.15).

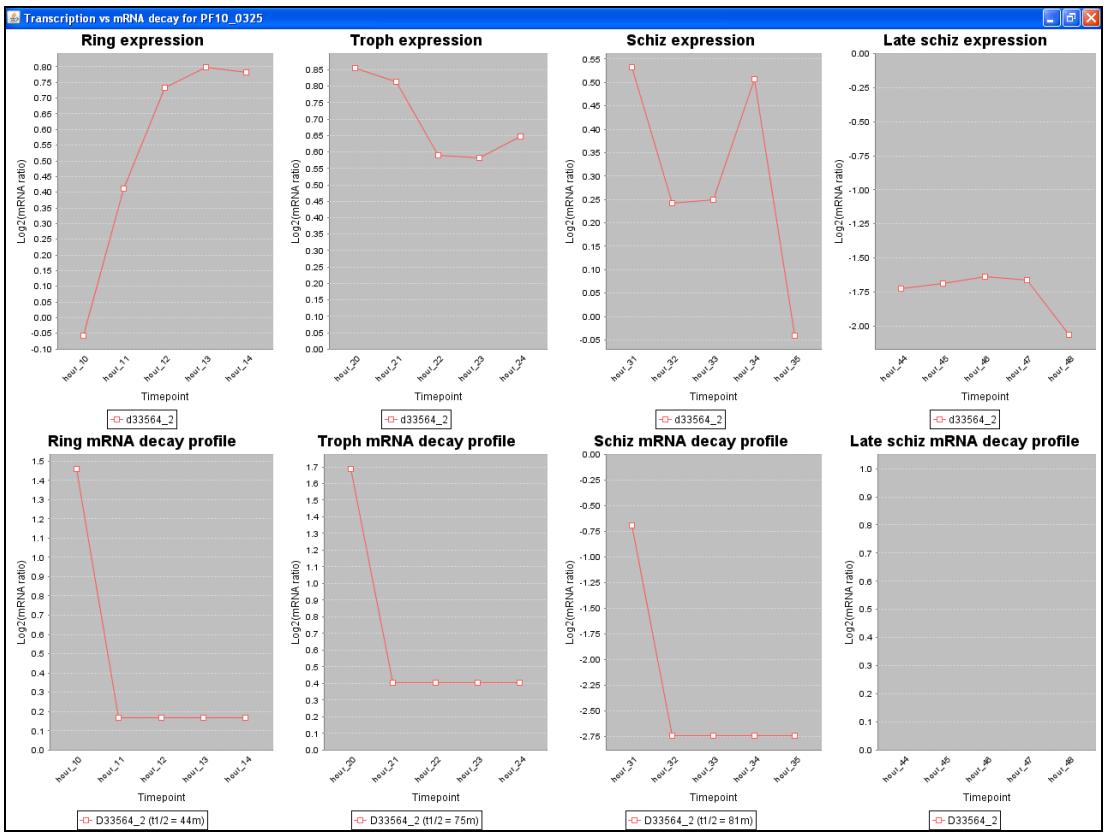


Figure 4.15. Screenshot of a set of mRNA abundance profiles (top row) versus the decay half life of the mRNA (bottom row) for the gene PF10_0325 at four stages of the IDC. Data is absent in the fourth mRNA decay profile because the recorded signal did not meet the quality control criteria applied by the authors of the study (Shock et al. 2007).

Finally, users can elect to display time-series data transformed to log base two, which helps when comparing expression profiles of genes with vastly differing abundances. It is also a necessary step for correct display of expression data from glass-slide arrays that represent ratios of mRNA expression (the datasets from Llinas et al. 2006). The $\log_2$ transformation step is always performed by the Java program before any expression data from these datasets are displayed (including in the Genome and Protein-Protein Interaction Viewers).

Access to quantitative data in the form of graphs is an important asset for expression data exploration because it allows users to check whether the trends they noticed in other viewers hold up under closer examination. User-selectable display of multiple genes on a single graph is an important feature that most other tools in the field currently do not offer.

Time-series graphs were implemented using a graph library called JFreeChart (JFreeChart; http://www.jfree.org/jfreechart/) as described in Section 4.1.2.2. A set of Java classes to display different types of expression data as graphs were developed; although specific details vary between classes, they are based around a common template for graph creation. For example, the SingleGeneTimeSeriesChart class is the 'parent' class for display of graphs for individual genes and is extended by 'child' classes SingleGeneOneRepTimeSeriesChart (for display of data with one profile per gene or protein) and SingleGeneMultiOligoTimeSeriesChart (for display of data with one profile per oligo). The parent class implements the methods *createChart*, which takes an instance of CategoryDataset (a JFreeChart class) as a parameter and returns an instance of JFreeChart (in this case a line chart), and *display*, which creates a window in which to display the graph. The child classes contain *createDataset* methods, which retrieve the expression data for a given gene in a chosen dataset and use it to create an instance of CategoryDataset. A similar format is followed for classes displaying times-series graphs for groups of genes (MultiGeneTimeSeriesChart, MultiGeneOneRepEachTimeSeriesChart and MultiGeneMultiOligoTimeSeriesChart), multiple datasets (MultiExperimentTimeSeriesChart) and mRNA decay data (SingleGeneMRNADecayCharts).

4.3.6.2 Mining expression data

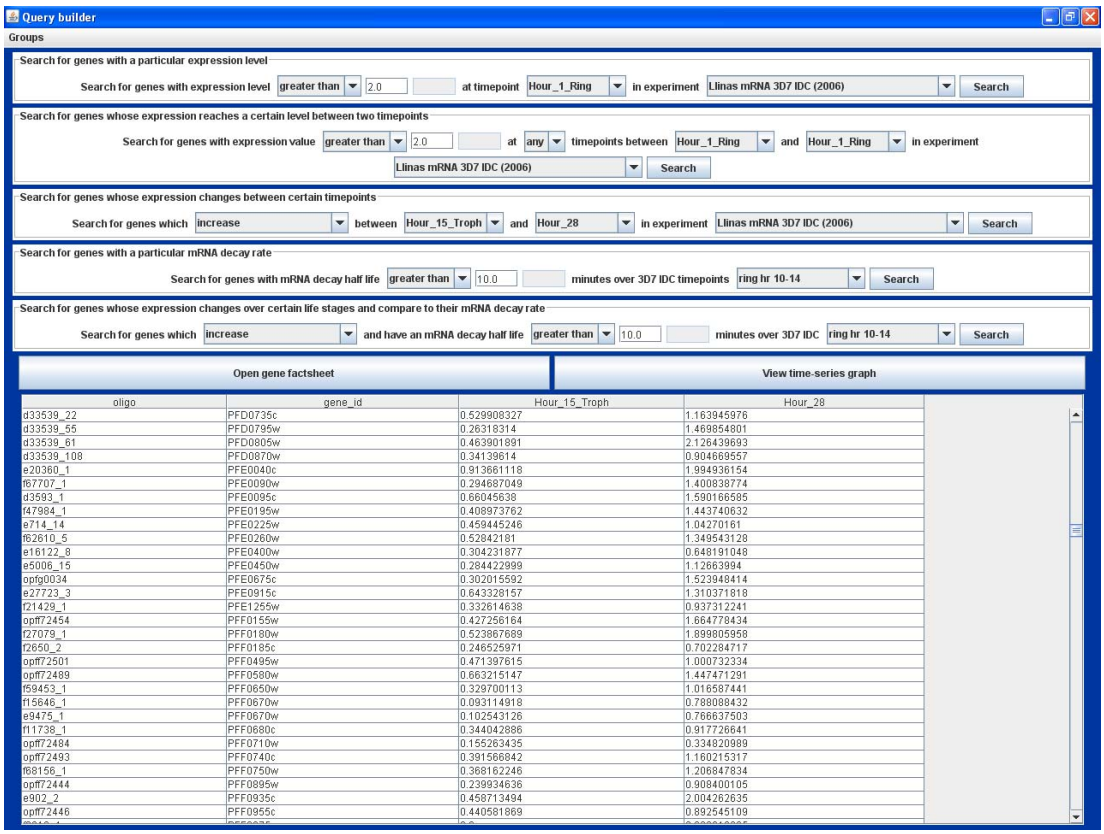
A Query Builder window accessible from the Expression Data Viewer aids users with mining expression data. Most other related tools do not offer users a dedicated, advanced search facility for expression data. The Query Builder window is coded for by an inner class within Transcriptome called QueryMenuHandler.

The Query Builder window provides users five 'template' queries, into which they can select options to build a query. The five options are:

1. Search for genes with expression above or below a certain cut-off value at one time-point;
2. Search for genes whose expression reaches or dips below a certain cut-off value during a range of time-points;
3. Search for genes whose expression significantly changes in a particular direction between two time-points;
4. Search for genes that have a certain mRNA decay half-life length;
5. Search for genes that could be being regulated, by combining options 3 and 4.

Following searches a data table is displayed (Figure 4.16), on which users can highlight genes to add to selected groups A or B. This functionality offers users a unique way of quantitatively mining expression data to find groups of genes that follow specific patterns of expression over a time-frame of interest. Properties of selected genes can then be visually investigated using the MaGnET Viewers.

Other options available from the data table are quick display of time-series graphs and gene fact sheets.



oligo	gene_id	Hour_15_Troph	Hour_28
d33539_22	PFD0735c	0.529908327	1.163945976
d33539_55	PFD0795w	0.26318314	1.469854801
d33539_61	PFD0805w	0.463901891	2.126439693
d33539_108	PFD0870w	0.34139614	0.904689557
e20360_1	PFE0040c	0.913661118	1.994936154
f67707_1	PFE0090w	0.294687049	1.400838774
d3593_1	PFE0095c	0.66045638	1.590166585
f47984_1	PFE0195w	0.408973762	1.443740632
e714_14	PFE0225w	0.459445246	1.04270161
f62610_5	PFE0260w	0.52842181	1.349543128
e16122_8	PFE0400w	0.304231877	0.648191048
e5006_15	PFE0450w	0.284422999	1.12663994
opfg0034	PFE0675c	0.302015592	1.523948414
e27723_3	PFE0915c	0.643328157	1.310371818
f21429_1	PFE1255w	0.332614638	0.937312241
opff72454	PFF0155w	0.427256164	1.664778434
f27079_1	PFF0180w	0.523867689	1.899805958
f2650_2	PFF0185c	0.246525971	0.702284717
opff72501	PFF0495w	0.471397615	1.000732334
opff72489	PFF0580w	0.663215147	1.447471291
f59453_1	PFF0650w	0.329700113	1.016587441
f15646_1	PFF0670w	0.093114918	0.788088432
e9475_1	PFF0670w	0.102543126	0.766637503
f11738_1	PFF0680c	0.344042886	0.917726641
opff72484	PFF0710w	0.155263435	0.334820989
opff72493	PFF0740c	0.391566842	1.160215317
f68156_1	PFF0750w	0.368162246	1.206847834
opff72444	PFF0895w	0.239934636	0.908400105
e902_2	PFF0935c	0.458713494	2.004262635
opff72446	PFF0955c	0.440581869	0.892545109

Figure 4.16. Screenshot of the Expression Data Viewer's Query Builder page. The panels at the top allow complex queries to be quickly constructed. Results are displayed in a table. By highlighting a particular gene or genes in the table, users can bring up expression profile graphs at the click of a button.

4.3.6.3 Attributes and methods of the Transcriptome class

Attributes of the Transcriptome class include the panels, buttons and text field required by the Expression Data Viewer, menu options to allow selection of datasets for graph display, and an instance of JFrame called 'queryWindow' to hold the Query Builder window.

Aside from the QueryMenuHandler class, inner classes defined within the Transcriptome class include a TSButtonHandler, which responds to user requests to

display time-series graphs by examining the selected datasets and number of entered gene identifiers to choose the correct type of graph to display.

4.3.7 Gene fact sheets and the Gene class

The Gene class represents a single gene. As such, it contains many attributes to describe a particular gene, including gene identifier, genomic location, exons, type, product name, aliases, gene and protein sequences, curated annotation, GO annotation, predicted InterPro domains and sequence features, structural data, interactions, orthologs and paralogs.

In addition, the Gene class contains a method to display a gene fact sheet that lists all the above information about a gene (*createGeneFactsheet*). The Gene class also contains several attributes that are required by the fact sheet, such as drawing panels. The fact sheet is designed around a series of panels, each displaying different types of information about the gene, organised into a number of tabbed panes within the fact sheet window (Figure 4.17). Furthermore, fact sheets provide useful link-outs to gene entry pages in other online resources, including PlasmoDB (http://www.plasmodb.org/plasmo/), GeneDB (http://www.genedb.org/), UCSC Malaria Genome Browser (http://areslab.ucsc.edu/cgi-bin/hgGateway), ApiCyc (http://apicyc.apidb.org/), DeRisi and Winzeler lab transcriptomic databases (http://malaria.ucsf.edu/comparison/index.php and http://chemlims.com/OPI20/MServlet.ChemInfo, respectively) and TDR Targets Database (http://tdrtargets.org/).

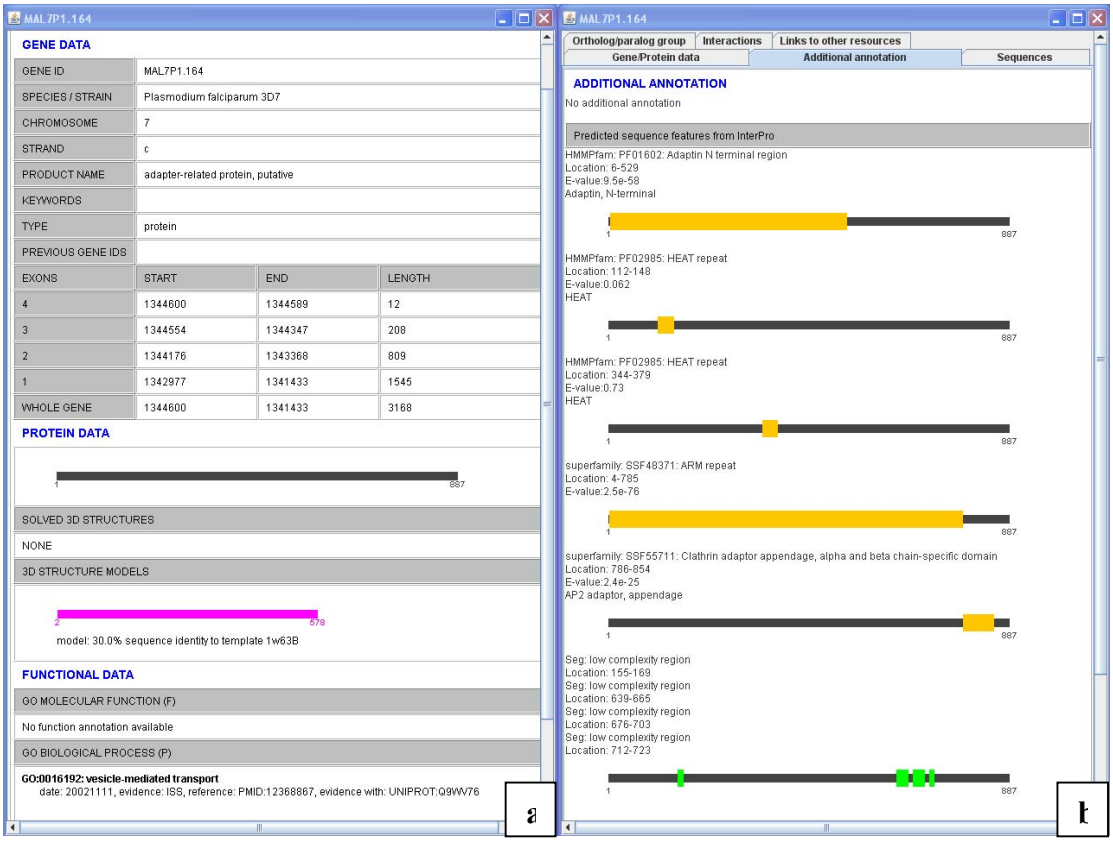


Figure 4.17. Screenshots of two pages from the fact sheet belonging to gene MAL7P1.164. The fact sheet incorporates data on genomic location, structure, function, interactions, orthologs and paralogs. Fact sheets also provide access to gene and protein sequences and link-outs to online databases, such as PlasmoDB (Bahl et al. 2003) and the UCSC Malaria Genome Browser (Chakrabarti et al. 2007). The first page (a) summarises the gene's genomic location, including a table of exons. Available protein structures are drawn as magenta bars below the region of the protein (black bar) they represent. GO annotations are listed in separate sections according to their ontology (function, process or component). The second page (b) presents curated annotation and predicted InterPro domains and sequence features, which are graphically represented by coloured blocks drawn over the corresponding region of the protein (black bar).

4.3.7.1 Protein structure visualisation

Figure 4.17a demonstrates display of protein structure information within the fact sheet. Protein structures are divided into experimentally solved and

comparatively modelled. Both are represented as magenta bars alongside the corresponding region of the protein. By double clicking on a structure, users may open the 3D coordinates of the structure in the Jmol molecular viewer (Jmol; http://www.jmol.org/) (Figure 4.18) (see Section 4.1.2.1 for details about the implementation).

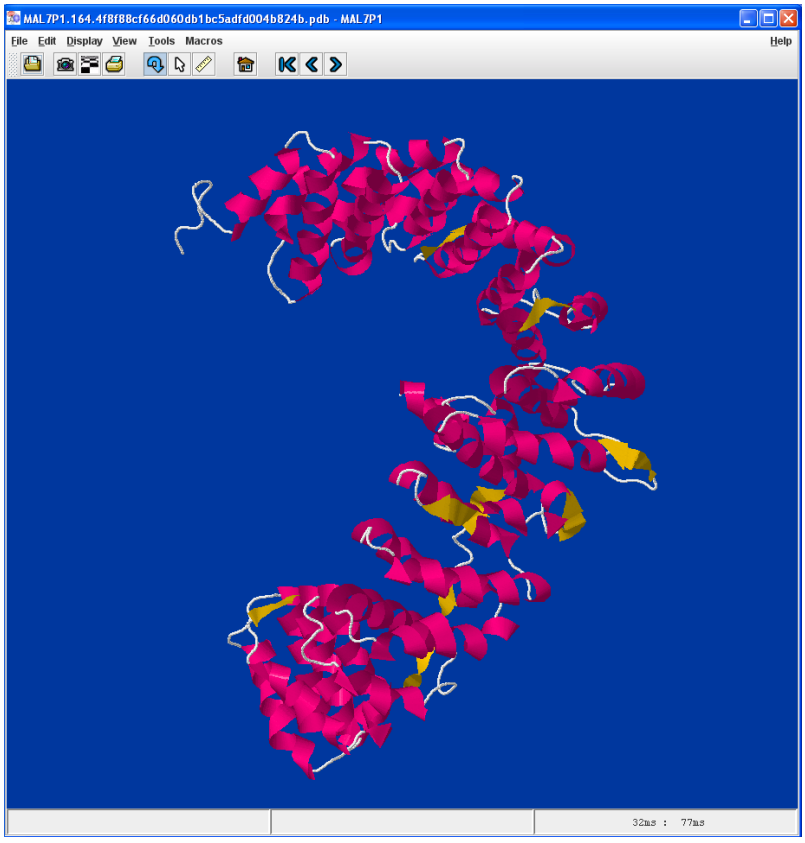


Figure 4.18. Screenshot of the modelled structure of the protein product of MAL7P1.164 displayed in the Jmol molecular viewing program (Jmol; http://www.jmol.org/). When Jmol opens, a custom script instructs it to display the structure as a cartoon representation of the secondary structure elements, with only the backbone residues displayed. Users can interact with Jmol through its menus and command console.

It was decided to include the Jmol application within the MaGnET software in order that users could easily visualise structures and immediately begin to interact with them without having to download their own molecular viewer. The complete version of Jmol is included (an alternative would have been to provide a basic, stripped-down version) so that users can make use of its full functionality, via its helpful menu system or RasMol-like commands (Bernstein 2000), with which many people are familiar.

This functionality provides a useful addition to MaGnET that is currently not implemented in most other tools in the field. For instance, the TDR Targets Database links to the model entry in ModBase but does not provide an integrated method for visualising protein structures, nor does it attempt to curate models to include only good qualities ones (TDR Targets Database; http://tdrtargets.org).

4.3.7.2 Attributes and methods of the Gene class

The main attributes and methods of the Gene class are described above. Other methods are defined that retrieve various data about a gene from the database and populate the corresponding attributes. For example, the methods *get3DModels* and *getPDBCodes* retrieve comparative models and PDB structure information for the protein. Inner classes such as ClickOnModelHandler and ClickOnPDBHandler respond to a user clicking on a magenta bar representing a comparative model or solved structure, respectively, by opening the corresponding structure in Jmol. Other inner classes are defined to represent particular data types or attributes of a gene, such as Domain (an InterPro predicted domain or sequence feature), SignalPeptide (a predicted signal sequence), Interaction (a single protein-protein interaction between the gene's product and another protein) and GO (a GO term annotation).

4.4 Availability

4.4.1 Online availability

MaGnET is freely and publicly available over the World Wide Web at `http://www.malariagenomeexplorer.org`. A Java applet version of MaGnET is available to use online via a web browser. Alternatively, users may download the application to their computer and run it using Java Web Start technology (Sun Microsystems; http://java.sun.com/products/javawebstart/). Java Web Start is included with the Java Runtime Environment (JRE) distribution. The Java Web Start application is preferred and encouraged above the applet because it has the advantage of significantly increased speed of use. Furthermore, once the application has been downloaded to the user's computer, subsequently they will be able to run it directly from their computer without requiring returning to the MaGnET website. At start-up, the application automatically checks for updates to the program and downloads them. Nevertheless, the applet version offers advantages where the user wishes to have instant access to MaGnET within their web browser or may not wish to download the application (for example, when using a public computer).

MaGnET runs satisfactorily on computers with the latest JRE for their operating system, including Windows XP with JRE 1.6, Mac OS X with JRE 1.5 and Linux with JRE 1.5. Furthermore, the MaGnET applet runs in the common web browsers, including Mozilla Firefox, Internet Explorer and Safari. Therefore, MaGnET should work on any system running JRE 1.5 or a later version; also, MaGnET should continue to work satisfactorily with future releases of Java, since they are backwards compatible.

The MaGnET database and website are currently hosted by the School of Engineering, University of California, Santa Cruz. It is hoped that a mirror server will soon be available at the University of Edinburgh, which will improve the speed of downloading the application, loading the applet and accessing the database for clients based outside the United States.

4.4.2 Downloadable version with database

A full, licensed version of the program and database, with source code, will be made available upon request. This will allow advanced users and developers to keep a local copy of the database, to which they could add their own data, such as gene annotations or expression data. Expansion of the database would, in most cases, require a slight modification of the Java code. Every effort has been made in the development of MaGnET to ensure that the program is robust and easily expandable to facilitate the addition of new datasets. It is also hoped that features of MaGnET will form the basis for new visualisation tools provided by the main malaria genome databases. For this purpose, access to the program source code will be essential. The use of discrete Java classes reflecting the organisation of the visualisation program means that it will be straightforward for complete, individual MaGnET viewers to be exported.

4.4.3 Documentation

The documentation available to MaGnET users includes a detailed set of help pages accessible from any point within the program. The help pages include helpful tips on how to make the most of the program's functionality as well as information about the datasets, including source and version details. An introductory overview of

the software and data (including screenshots) is provided on the MaGnET website for new users. The website also includes a useful tutorial which is recommended for all first-time users.

4.4.4 Security considerations

In order to allow users to read and write files of gene identifiers MaGnET needs to have special permission to access their local file system. The necessary permission can granted by way of a security certificate digitally signed by the program developer that is included with the applet or application JAR (program package). The MaGnET applet and application JARs have been signed so that they can access the user file system; the first time a user accesses the program they must accept the security certificate. Behaviour of the browser can vary in this situation according to the individual settings of the user.

Aside from the reading and writing of files, MaGnET currently resembles a "closed" system. Other websites and programs are currently not allowed to link into the program, because this opens up a lot of technical issues about security and session caching which would need to be dealt with in order to make MaGnET fully "online" and interactive. To achieve this was not possible within the scope of this project, although it would be useful in future for other programs to be able to open MaGnET with particular a set of genes pre-selected, perhaps in a certain viewer. Nonetheless, within the scope of the permissions granted to MaGnET it is able to open a web browser and direct the user to a particular web address, which meant that link-outs to online databases could be included.

4.5 Discussion

The Malaria Genome Exploration Tool provides advanced visualisation for a range of *P. falciparum* functional genomic data. It encourages users to explore the available annotation by providing simplified search facilities that allow them to rapidly locate genes of interest. The inclusion of several linked data viewers offers the user a choice of entry points and allows them to follow their analysis down any pathway. Users may browse protein-protein interactions by navigating the Protein-Protein Interaction Viewer, where they can quickly pull up a network of interactions for a protein of interest. Using the Genome Viewer and associated Chromosome Viewer allows genes to be quickly located and their genomic context to be examined at various levels of resolution. The Expression Data Viewer allows the biological context of gene and protein activity to be explored, at the level of individual genes and small groups. A comprehensive search interface allows the user to retrieve sets of genes following a similar pattern of expression at a particular stage of the life cycle. Users may quickly search the database for genes and annotations that match keywords or analyse a list of annotation summaries for a selected group of genes using the Data Analysis Viewer.

MaGnET offers advantages over existing *Plasmodium* genome databases (such as in Table 1.5) by allowing users to select and manipulate groups of genes and to carry these groups forward across viewers. This approach makes it simple for users to explore the features in common for a set of genes, and to discover previously unknown relationships. Meanwhile, MaGnET makes available detailed annotation for individual genes in the form of helpful fact sheets that can be accessed at the

click of a button. The data is separated on tabbed "pages" to facilitate rapid navigation to the annotation of interest.

MaGnET also features a novel method for visualising time-series gene and protein expression data mapped onto either genomic location or protein-protein interaction networks. This facility opens up a new dimension for functional genomics data exploration by allowing malaria biologists for the first time to use *in silico* tools to generate hypotheses about how genes work together to achieve their function. For instance, where co-expressed genes occur in the same region of a chromosome it suggests that they may be subject to the same transcriptional control mechanism and may be involved in the same cellular process. Using this principle of "guilt by association", users can come up with new hypotheses about the possible function of previously uncharacterised genes. MaGnET can be used to explore all aspects of a protein of interest in theory before deciding on follow-up experiments to perform, either computationally or in the laboratory.

Integration of different data-types can be also used to discover erroneous annotation. For instance, if the expression data indicates that proteins which are reported to interact in yeast two-hybrid experiments are never expressed during the same life cycle stage, then it may be assumed that the interaction is not likely to occur in a biological context. Another example might be when a gene has been annotated (computationally) with a particular function, but there is no experimental evidence and it can be closely associated with a group of genes with a different function, then this indicates that the original annotation is likely to be false.

By selectively including relevant functional genomics data, MaGnET eliminates a lot of the "clutter" that is inherent in many of the currently available

resources for *Plasmodium* data. This serves to minimise the sensation of "data overload" that is sometimes felt by biologists seeking to use bioinformatics tools and genome databases. Careful filtering out of low-quality annotation means that users can be confident when using MaGnET that they will only be presented with useful and reliable annotation. In this sense, MaGnET helps users by making some of the decisions they face when using other data resources for them.

4.5.1 Comparison of MaGnET to similar tools

When MaGnET was designed in late 2004, there were still very few tools around that implemented the functionality that MaGnET set out to achieve (see Chapter 2 – System Design). As Tables 1.6, 1.7 and 1.8 demonstrate, the majority of similar tools focus on visualisation of one major type of data, such as gene expression data [GenePilot (TG Services; El Sobrante, CA, USA), GeneXplorer (Rees et al. 2004), TableView (Johnson et al. 2003)], cellular pathways [PathwayTools (Karp et al. 2002), PathwayExplorer (Mlecnik et al. 2005), SHARKview (Hyland et al. 2006)] or networks [Osprey (Breitkreutz et al. 2003), Bioverse (McDermott et al. 2005)]. Recently, the designers of these tools have recognised the importance of giving users access to annotation data that will help with their analysis, and have done so by including GO and other annotation, as well as link-outs to online databases. Furthermore, integrated visualisation of gene expression data in the context of pathways and networks has become commonplace. PathwayTools, PathwayExplorer, Bioverse, SHARKview, Yeast Exploration Tool Integrator (YETI) (Orton et al. 2004) and GenMAPP (Salomonis et al. 2007) all offer pathway or network visualisation that uses a colour code to display the expression level of the components. Additionally, PathwayTools has recently implemented

integrated visualisation of expression data within genomic maps, as well as providing a simple step-through mechanism using arrow buttons for displaying changes in expression over a time-course experiment.

The precedent set by YETI (Orton et al. 2004) and MaGnET for linking multiple data viewers to create a multi-faceted "workbench"-style tool has been echoed in the development of other software. As mentioned, PathwayTools is one of the first to offer visualisation tools for browsing both cellular pathways and genomic location. GenMAPP is also developing a range of viewers for integrating various data types, including pathways, expression data and SNPs. TM4's MultiExperiment Viewer (MeV) (Saeed et al. 2006) is currently developing modules for viewing metabolic pathways and genome maps with expression data overlaid to augment its existing microarray visualisation tools.

Despite the recent emergence of generic tools offering integrated visualisation of functional genomics data, few of them offer the range of data that MaGnET does. One data type lacking from all the resources listed in Tables 1.6, 1.7 and 1.8 is protein tertiary structure information. Since protein structure ultimately determines function, this information adds another dimension to a protein's annotation. Even though there are few experimental structures available for *Plasmodium* proteins, comparatively modelled structures are available for approximately a third of the proteome. Providing a simple way to access this information is a valuable contribution of MaGnET to the field.

4.5.1.1 Comparison of MaGnET to related Plasmodium-focussed tools, including detailed comparison to the Plasmodium Genome Resource, PlasmoDB

MaGnET was developed to plug a perceived gap in the market for a tool to facilitate malaria researchers wishing to explore the growing amount of functional genomic data available for *Plasmodium*. There are still few tools available offering all the capabilities provided by MaGnET. The tools discussed in the above section are mainly generic tools that are not specifically set-up to support *Plasmodium* data and where they do (e.g. PathwayTools) the data is moulded into a standard framework that does not recognise the unique properties and special requirements of this data, as discussed in Chapter 1.

The Plasmodium genome databases described in Table 1.5 – PlasmoDB (Bahl et al. 2003), GeneDB (Hertz-Fowler et al. 2004), NCBI Malaria Genetics and Genomics (http://www.ncbi.nlm.nih.gov/projects/Malaria/), WHO/TDR Malaria Database (http://www.wehi.edu.au/MalDB-www/who.html), Broad Institute *P. falciparum* Database (http://www.broad.mit.edu/annotation/genome/plasmodium_falciparum_spp/MultiHome.html), and TIGR Parasites Databases (http://www.tigr.org/tdb/parasites/) – are excellent resources for researchers wishing to find out what is already known about a protein they are interested in. However, few of them are dedicated to providing tools for the mining of functional genomic datasets to discover new trends and compare features of groups of genes.

The remaining *Plasmodium* resources in Table 1.5 are bioinformatic tools for making new annotations about *Plasmodium* genes. They tend to focus on a specific

aspect of gene structure or function; for instance, MalariaBase (http://malariabase.org/) is a tool for predicting new gene functions based on BLAST (Altschul et al. 1990) results and analysis of sequence motifs. The UCSC Malaria Genome Browser (Chakrabarti et al. 2007) was designed to facilitate annotation of gene structure, which it does by combining evidence of transcribed and translated sequences, automated exon predictions and comparative genomics. Finally, the SAMP database (part of the MalPort server) (Joubert and Joubert 2008), combines various predictions of protein sequence features and structural data to generate lists of priority targets for further structural characterisation. All of these tools will be very useful for their own specific purposes; however, none of them covers the range of data or the remit of MaGnET: providing visualisation of integrated functional genomic data.

A recent development in the field, MADIBA (also part of the MalPort server) (Law et al. 2008), provides malaria researchers with a tool for the analysis of clusters of co-expressed genes. Whilst the purpose and output of MADIBA are clearly very different to MaGnET, some of the themes running through the MaGnET design are recognisable in MADIBA. For instance, MADIBA also encourages users to explore common features of a group of genes by providing tools to analyse chromosomal location and over-representation of GO terms in gene lists. MADIBA provides bare-bones visualisation of results, although this is limited to locations of the clustered genes in KEGG pathways (a topic not included in MaGnET), display of over-represented GO terms in a hierarchical tree and basic visualisation of chromosomes with gene locations marked. Despite some similarities, MADIBA lacks some of the main features of MaGnET. For example, MADIBA does not store many of the

functional genomic datasets that MaGnET does, including expression data and protein-protein interactions. Therefore, users cannot discover new clusters or groups of genes via browsing the available datasets as they can when using MaGnET; they must already have a pre-defined list of genes to use MADIBA. Furthermore, MADIBA does not integrate different datasets like MaGnET, so different aspects of gene function can only be examined individually. This would make it very difficult to explore relationships between different data-types; for example, it would be impossible to view expression patterns across genomic regions or interaction networks using MADIBA.

Nevertheless, PlasmoDB (Bahl et al. 2003) remains the primary information resource for malaria researchers working on all aspects of *Plasmodium* biology. Table 4.2 provides an overview of the data content of MaGnET as compared to PlasmoDB and shows that most of the main data types are shared between the two. PlasmoDB is foremost a genome database, so its visualisation tools are focussed around the genome browser, useful features of which include customisable track display. Although MaGnET provides a Genome Viewer, its purpose is for use in conjunction with other functional genomic data; for example, display of gene groups, paralogs and transcription patterns in a genomic context. In other cases, MaGnET offers far more advanced visualisation, such as a Protein-Protein Interaction Viewer and the ability to easily compare expression profiles of multiple genes. MaGnET does not feature all possible types of expression data that are included within PlasmoDB, such as ESTs and wild type (WT) versus ligands for merozoite invasion of erythrocytes knock-out (KO) studies, because the development of MaGnET

focussed on visualisation of time-course data. However, MaGnET could be straightforwardly adjusted to incorporate other types of expression data.

Data access	PlasmoDB	MaGnET
Gene location	Standard Genome Browser (GBrowse) – zoom-able, limited to sections of chromosome	Genome Viewer/Chromosome Viewer – zoom-able, up to whole chromosome or whole genome view
Sequences	FASTA protein/DNA	FASTA protein/DNA
Gene Ontology	Search terms by keyword. Listed on gene page.	Search terms by keyword. Listed on gene page. Compare GO for multiple genes in Data Analysis Viewer.
Results of other function prediction tools	InterPro predicted sequence features. Listed on gene page. Search by keyword.	InterPro predicted sequence features. Visualised on fact sheet. Search by keyword. Compare annotation for multiple genes in Data Analysis Viewer.
Protein interactions	Table of interactions on gene page (one yeast two-hybrid study).	Table of interactions on gene fact sheet. Interactive maps in Protein-Protein Interaction Viewer (one yeast two-hybrid study).
Expression data	Time-series expression profile graphs for single genes. MRNA data from three time-course studies. WT v KO expression data for erythrocyte invasion pathway.	Time-series expression profile graphs for single genes/proteins and for small groups. MRNA and protein data from five time-course studies. Extensively searchable via Expression Data Query Builder.
Metabolic pathways	Links to PlasmoCyc and Malaria Metabolic Pathway DB	Links to PlasmoCyc
Literature	Link out to Malaria Literature DB	Link out to Malaria Literature DB
Protein Structures	Link out to PDB and limited set of models (Gowthaman et al. 2005).	Locally stored PDB and modelled structures. Large set of high quality models from ModBase.
Orthologs/paralogs	List of orthologs in other *Plasmodium* genomes and in-genome paralogs. Link out to OrthoMCL DB from gene page.	List of orthologs in several *Plasmodium* genomes and in-genome paralogs. Link outs to OrthoMCL DB and orthologs in PlasmoDB from fact sheet.
ESTs	EST libraries from several species/strains.	--
SNPs	SNPs for several *P. falciparum* strains.	--

Table 4.2. Overview comparison of MaGnET data content to PlasmoDB.

Table 4.3 provides an overview of the main interface functionality of MaGnET compared to PlasmoDB. In PlasmoDB's role as main information provider

for the malaria research community, it seeks to provide a comprehensive set of search tools to facilitate data mining. While it is obviously necessary to be able to search the MaGnET database and select groups of genes to transfer to other viewers, it is not the primary purpose of MaGnET, so relatively modest search functionality is implemented.

PlasmoDB's data display is mainly organised around single gene pages listing all available information about a gene. There is limited functionality for comparing features of groups of genes. However, MaGnET encourages the use of gene groups for discovery of patterns and features in common across subsets of genes. MaGnET provides two "place-holders" for groups of unlimited size, which are assigned unique colours that are maintained when groups are transferred between different data viewers. MaGnET's ability to view a set of genes in the context of different datasets is a useful feature not currently implemented by PlasmoDB. That MaGnET has the option to hold two groups at once affords an additional level of functionality, because the groups' features can easily be compared and common genes discovered. Moreover, other tools in the field that do facilitate manipulation of groups of genes only implement a single group at a time approach. For example, the MADIBA toolkit for microarray cluster analysis (Law et al. 2008) allows users to upload a select set of genes and examine various properties of the group (see Table 1.5). By providing users with the option to select two sets of genes at a time MaGnET opens up an entirely new spectrum of possibilities for cross-cluster comparisons. In many cases users will be interested in the intersection between the two groups; for example, users could select groups of genes that are being up-regulated during a particular life cycle stage in two different microarray experiments

and then examine the genes that are present in both groups using this technique. When formulating hypotheses based on exploratory analysis, the more lines of evidence that can be drawn from the stronger and more focussed the hypothesis can be.

Features	PlasmoDB	MaGnET
Purpose	Database/data warehouse	Workbench/visualisation tool
Display organisation	Primarily single gene focussed. Gene neighbourhood available in genome browser.	Range of displays from whole genome down to small groups and single genes.
Selection of multiple genes	Combine search results to create a gene list.	Select and modify two sets of genes, carried forward between displays.
Compare genomes	View syntenic regions between certain genome pairs.	--
In-genome paralog view	--	View genomic location of paralogs. Easily add paralogs to a group.
User customisable	Select tracks to display in genome browser. Upload user data as custom tracks.	Select colours used for visualisation. Downloadable version could be modified to include user data.
Searchable	Many search options. Simple interface.	Several search options, especially for expression data. Can search using ID and keyword lists.
Integrated	Can move between genome browser and gene pages. Customisable track display on genome browser can display, for example, ESTs.	Fully integrated. Carry gene selections between displays. Display expression data mapped onto genome location or protein interaction map. Gene fact sheets available at any point.
External links to other resources	Many, including literature, orthologs, drug targets, UCSC genome browser	Many of the same
Other tools	BLAST, downloadable files, subcellular localisation prediction	Visualisation of protein structures with integrated Jmol viewer.
Store results	Save file with selected data about genes resulting from search.	Save file with IDs of selected set of genes.
Size	Large-scale, comprehensive, consortium.	Lightweight, portable, relevant data only, one group product.
Future lookout	Need to keep up with inclusion of new data leaves little time for developing advanced visualisation/exploration tools.	Framework in place, many possible directions for future expansion, e.g. visualisation of pathways and comparative genomics data.

Table 4.3. Overview comparison of MaGnET interface functionality to PlasmoDB.

4.5.2 Future improvement to and expansion of MaGnET

There are many directions in which the development of MaGnET could be taken. The most interesting and useful possibilities are summarised below.

1. Comparative Genomics Viewer

Comparative genomics has much to offer for investigating *Plasmodium* biology, including host specificity, drug resistance, parasite evolution and species and strain specific behaviours, such as rosetting. Therefore, it would be useful to extend the MaGnET Genome and Chromosome Viewers to facilitate comparative analyses between two or more *Plasmodium* genomes. The proposed Comparative Genomics Viewer would allow genomes of the user's choice to be automatically aligned and regions of synteny to be mapped. The existing functionality for viewing the genomic location of gene families in *P. falciparum* could be extended for all available genomes. Comparative genomics information will also allow functional annotations made in one species/strain to be transferred to another; for example, protein interaction networks may be assumed to be largely conserved, at least between isolates. At the gene/protein level, it will be useful to be able to compare the predicted structures and domains and to see the location of SNPs, insertions and deletions.

2. Pathway Viewer

Providing a Pathway Viewer would be a useful way to enhance the proteome viewing capabilities of MaGnET currently provided by the Protein-Protein Interaction Viewer. Much third-party software already exists for Pathway Viewers, so it may be possible to modify some open-source Java code to create a simple

Pathway Viewer. Alternatively, the MaGnET Protein-Protein Interaction Viewer itself could be modified, which would require the inclusion of non-protein components, such as chemical cofactors and substrates.

The main problem lies in obtaining *Plasmodium* pathway data. The Malaria Metabolic Pathways Database (Ginsburg 2006) is a comprehensive set of manually-curated metabolic and cellular pathways, but the annotation and gene associations are not straightforward to retrieve (there are no downloadable text files containing the information). The PlasmoCyc pathway database (Yeh et al. 2004), despite providing downloadable annotation files, is vastly incomplete and out-of-date [the ApiCyc *P. falciparum* pathway database (http://apicyc.apidb.org/) is currently just another incarnation of PlasmoCyc, but now that it is part of the PlasmoDB family it is more likely that it may be updated in future].

3. Whole-genome scale visualisation of expression data

In order to fully explore time-series expression data to discover patterns involving groups of genes with changing expression under particular conditions, a novel tool providing whole-dataset scale visualisation is needed. During the course of MaGnET's development, time was spent investigating the possibility of modifying an existing tool to provide this functionality within MaGnET. A collaboration with Professor Jessie Kennedy at Napier University, Edinburgh, was initiated with the aim of incorporating a modified version of a software tool for animated visualisation of microarray time-course data: Time-series Explorer (TSExplorer) (Craig et al. 2005). TSExplorer provides animated scatter-plot views of expression data, which allows the user to hone in on subtle changes involving groups of genes following the same pattern of expression over a short time-frame. Unfortunately, due to a number of

circumstances, it was not possible to complete the integration of TSExplorer with MaGnET in the time-frame of this PhD work. It is hoped that a future version of MaGnET will include the functionality provided by TSExplorer, which will greatly enhance the ability of malaria biologists to explore time-course expression data.

4. Comparison of expression profiles across datasets

The existing functionality within MaGnET to compare individual gene expression profiles across multiple datasets (useful for reinforcing confidence in the measurements) is very limited. It would be useful to provide an improved method for comparing at least two datasets, which would require the corresponding time-points to be accurately mapped. Since the currently available *Plasmodium* expression datasets are very different, it is difficult to determine exactly corresponding time-points. Hopefully, as more experiments are completed, it will become easier to find similar datasets to compare.

5. List of orthologues or counterpart proteins in model organisms and comparison of protein-protein interaction networks

During discussions with biologists an often requested feature was the mapping of *P. falciparum* gene names to their orthologues or functional counterpart proteins in model organisms, such as mouse and yeast. Another level to this would be to include functional information, such as protein-protein interaction data, from the model organisms and allow users to compare it with the *P. falciparum* data.

6. Improved Data Analysis Viewer

The Data Analysis Viewer could be improved by adding more search options and developing a Query Builder similar to that in the Expression Data Viewer to easily construct advanced queries.

The layout of the data table that is returned when a user requests a summary of the genes in their selected groups can be improved. The table could be enhanced by making it fully interactive, so that users can decide exactly what information is displayed and for which genes.

7. Additional protein structure information

The information provided about protein structures could be usefully supplemented with structural classifications from SCOP (Structural Classification of Proteins) (Murzin et al. 1995) or CATH [Class (C), Architecture (A), Topology (T) and Homologous superfamily (H)] (Orengo et al. 1997).

A novel use of the protein structure data would be to map the location of sequence variations between isolates onto the solved or predicted structure. This information can indicate areas of the protein that are important for function and changes in sequence and structure that might be causing an observed phenotypic difference.

8. Further link-outs

The gene fact sheet could be enhanced by providing links to descriptions of GO and InterPro terms (adding a link to the relevant resource webpage describing the term would be the most convenient way to do this). It would be useful to include link-outs to other online databases; for example, to relevant literature.

9. Automatic updating of database

The MaGnET database currently needs to be updated manually by downloading the relevant text files from their sources. It would be useful to develop an automated workflow to check for updates, retrieve the files from the web and run the update scripts without requiring manual intervention. While this would be possible to implement for certain datasets, such as gene annotation files, it would not be possible for others, such as expression datasets, which are downloaded once as publication supplementary data files. The addition of a new expression dataset requires changes to be made to the Java code, due to the customisation necessary for display of different types of data. If new expression datasets could be added to MaGnET without changing the Java code this would be a useful improvement, but would probably require major changes to the way in which expression data is read in and stored in the database.

ACKNOWLEDGEMENTS

Some visual aspects of the MaGnET user interface design were based on work by Richard Orton (Orton 2006) but were re-programmed for MaGnET's specifications. Part of the MaGnET Protein-Protein Interaction Viewer Java class extends initial work by Richard Orton (Orton 2006) (see Section 4.3.5 for further details).

5. DEMONSTRATION OF MAGNET EXPLORATION

Overview

This chapter is devoted to demonstrating that MaGnET may be used to illustrate the results of several recently published studies about gene function. The first "mini-study" shows that MaGnET effectively displays the majority of results reported in the publication of a gene expression study. The experimental dataset from this publication is included in the MaGnET database, so it has significantly influenced MaGnET design. Two other mini-studies describe research from a combination of bioinformatic and experimental investigations into the functions of two novel genes families. These studies also demonstrate that MaGnET analysis can reproduce a similar pathway to that used by the authors involving alternative bioinformatic tools to achieve the same result. The latter studies did not influence MaGnET design in any way.

The publications also included some experimental work that could not be exactly reproduced using MaGnET; however, where possible, related data available within MaGnET is consulted to demonstrate their findings. In a few places additional discoveries were made that were not recorded in the original publication.

Section 5.4 will evaluate MaGnET's strengths and weaknesses when visualising the results described in these publications. The advantages and limitations of MaGnET over other publicly-available tools to explore the data both used by and reported in the publications will also be discussed. All figures included for illustration are MaGnET screenshots.

The subsequent chapter (Chapter 6) will build upon this foundation and aim to demonstrate that MaGnET can be used to discover previously undocumented trends and generate hypotheses about gene function.

5.1 Gene expression profiling of the Intraerythrocytic Developmental Cycle (Llinas et al. 2006)

The first in-depth, genome-scale study of *P. falciparum* gene expression during the Intraerythrocytic Development Cycle (IDC) was conducted by Llinas et al. (2006). In this paper the authors compared microarray gene expression profiles recorded for nearly all genes in the laboratory strains 3D7 and Dd2 with results from an earlier microarray study of the HB3 laboratory strain (Bozdech et al. 2003). For a description of the array used in these studies refer to Section 1.5.1.2. Here, select results of the Llinas et al. 2006 study will be demonstrated using various features of MaGnET. It should be noted though that the data stored in the MaGnET database does not represent absolute values of expression level, but instead relative expression ratios. From such ratios it is not possible to determine the absolute level of expression of individual oligonucleotides. Absolute expression values were not made publicly available by the authors of the studies; therefore, the assertions made in the Llinas et al. paper using absolute expression values cannot be reproduced here.

5.1.1 Variability in gene expression

For the majority of genes timing of expression during the IDC was strongly correlated between strains 3D7, HB3 and Dd2 (Llinas et al. 2006). However, the few exceptions occurred mainly for genes encoding surface proteins involved in antigen

presentation. For example, the mRNA for gene MAL13P1.344 (a probable ATP-binding cassette transporter) is clearly differentially expressed in HB3 compared to 3D7 and Dd2 (Figure 5.1).

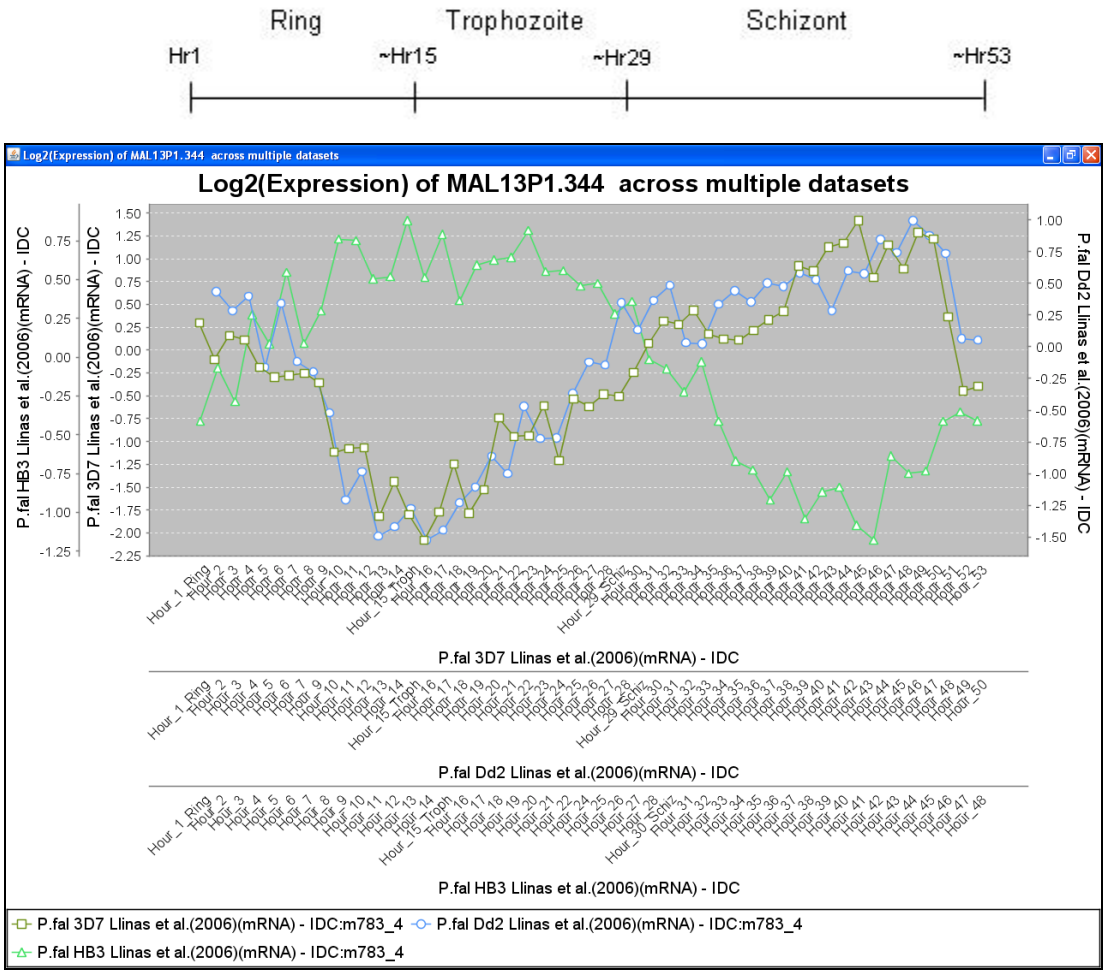
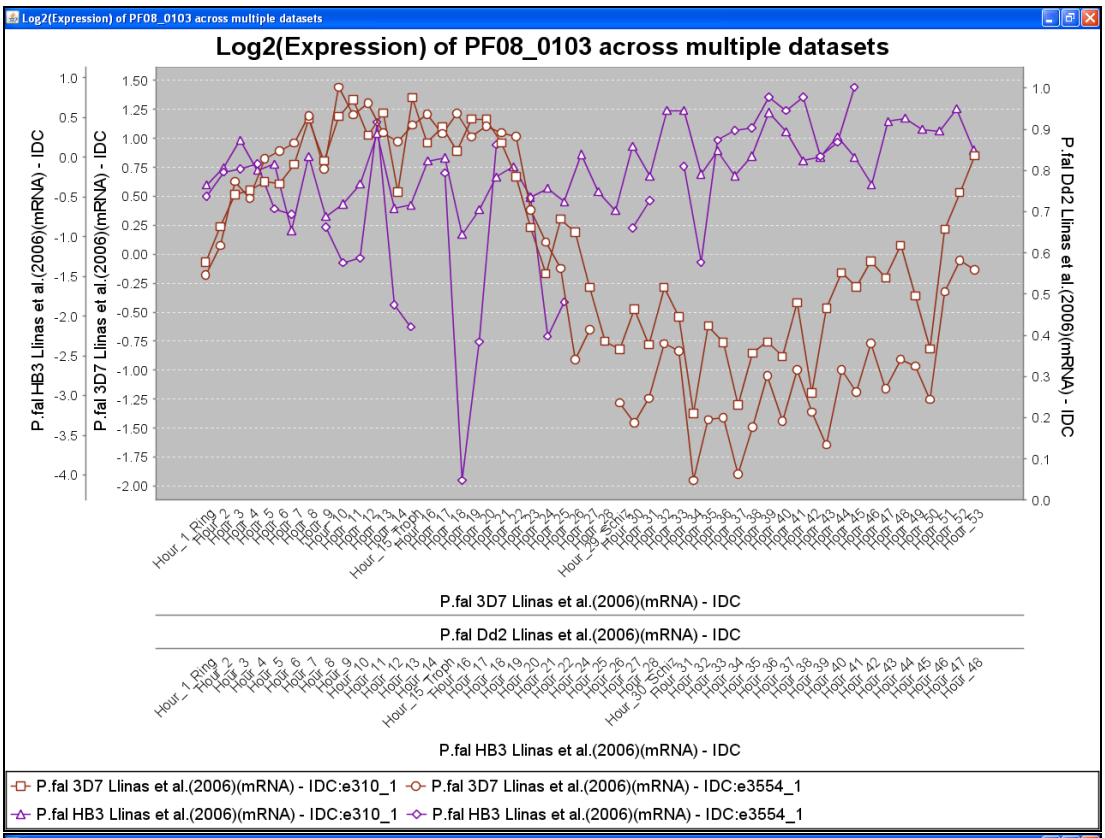


Figure 5.1. Time-series expression profiles for ATP-binding cassette transporter-encoding gene MAL13P1.344 in the 3D7, Dd2 and HB3 strains. Expression profiles in the 3D7 and Dd2 strains are very similar, but HB3 expression is almost opposite. At the top of the figure is a bar showing the timing of life cycle stages during the IDC for 3D7 strain parasites. The timings of life cycle stages for Dd2 and HB3 strain parasites are similar to those for 3D7, but the overall length of the IDC is approximately three and five hours shorter, respectively. Note that data are missing for hours 23 and 29 of the HB3 IDC. The information is included here for reference and also applies to all other mRNA expression profile graphs from the Llinas et al. 2006 dataset.

The *P. falciparum* erythrocyte membrane protein 1 (*Pf*EMP1)-encoding genes PF08_0103 and PFB0010w show a large variation in expression between strains 3D7 and HB3 (Figure 5.2). For both genes data are missing for Dd2; oligos which had missing values for greater than 40% of time-points were excluded from the MaGnET database (and also from the original analysis) (Llinas et al. 2006). Aside from technical issues, an oligo not being detected in one strain must be due to either transcriptional differences or significant deletions or polymorphisms. Indeed, PFB0010w lies in a region known to be silenced or deleted in Dd2 parasites (see Section 5.1.2).

The S-antigen (encoded by gene PF10_0343) is located in a known polymorphic region. The expression of this gene is not detectable in Dd2 or HB3 because the representative oligo lies within a highly diverse region of the sequence (Figure 5.3) (Llinas et al. 2006)

.


Log2(Expression) of PF08_0103 across multiple datasets
P.fal HB3 Llinas et al.(2006)(mRNA) - IDC
P.fal 3D7 Llinas et al.(2006)(mRNA) - IDC
P.fal Dd2 Llinas et al.(2006)(mRNA) - IDC
P.fal 3D7 Llinas et al.(2006)(mRNA) - IDC:e310_1
P.fal 3D7 Llinas et al.(2006)(mRNA) - IDC:e3554_1
P.fal HB3 Llinas et al.(2006)(mRNA) - IDC:e310_1
P.fal HB3 Llinas et al.(2006)(mRNA) - IDC:e3554_1

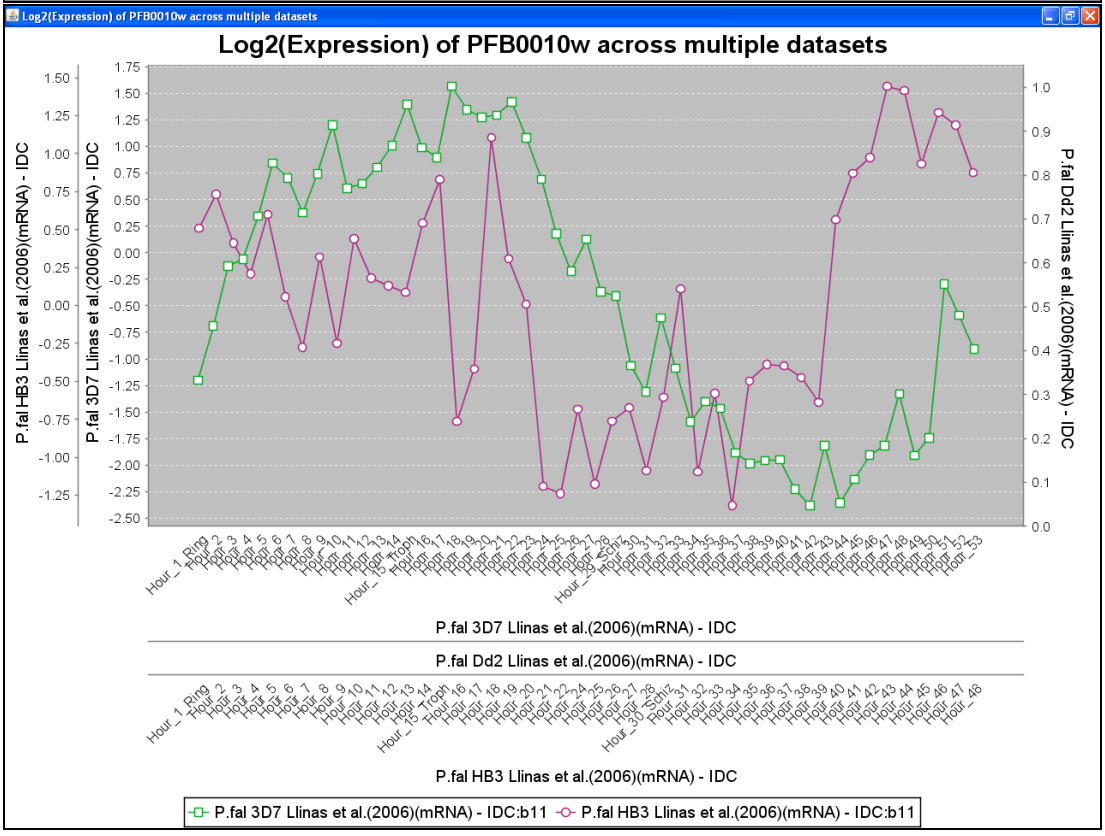

Log2(Expression) of PFB0010w across multiple datasets
P.fal HB3 Llinas et al.(2006)(mRNA) - IDC
P.fal 3D7 Llinas et al.(2006)(mRNA) - IDC
P.fal Dd2 Llinas et al.(2006)(mRNA) - IDC
P.fal 3D7 Llinas et al.(2006)(mRNA) - IDC:b11
P.fal HB3 Llinas et al.(2006)(mRNA) - IDC:b11

Figure 5.2. Time-series expression profiles for *Pf*EMP1-encoding genes PF08_0103 (top panel) and PFB0010w (bottom panel) in the 3D7 and HB3 strains (no data are available for Dd2). The expression profiles of both genes vary considerably between the two strains.

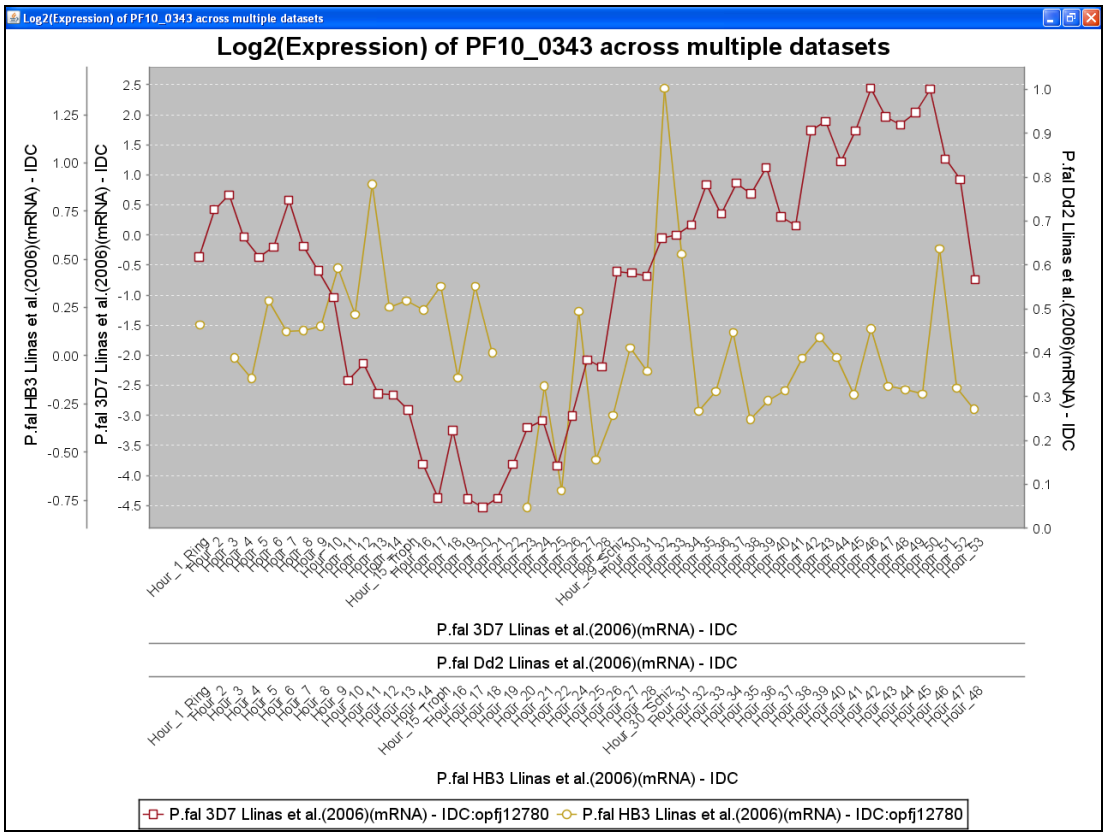


Figure 5.3. Time-series expression profiles for the S-antigen-encoding gene PF10_0343 in the 3D7 and HB3 strains (no data are available for Dd2). High sequence polymorphism in the region represented by this probe caused a lack of observed expression in HB3 and Dd2.

Llinas et al. report that the gene (PF11_0512) encoding ring-infected erythrocyte surface antigen 2 (RESA-2) demonstrates variation in expression between strains. Figure 5.4 shows that although expression profiles appear similar between strains HB3 and Dd2, there are differences to 3D7 parasites. The expression profiles of the two 3D7 oligos vary widely, which could be an indicator of sequence

variation, or possibly alternative splicing. Indeed, 3D7 is missing data for a third oligo altogether.

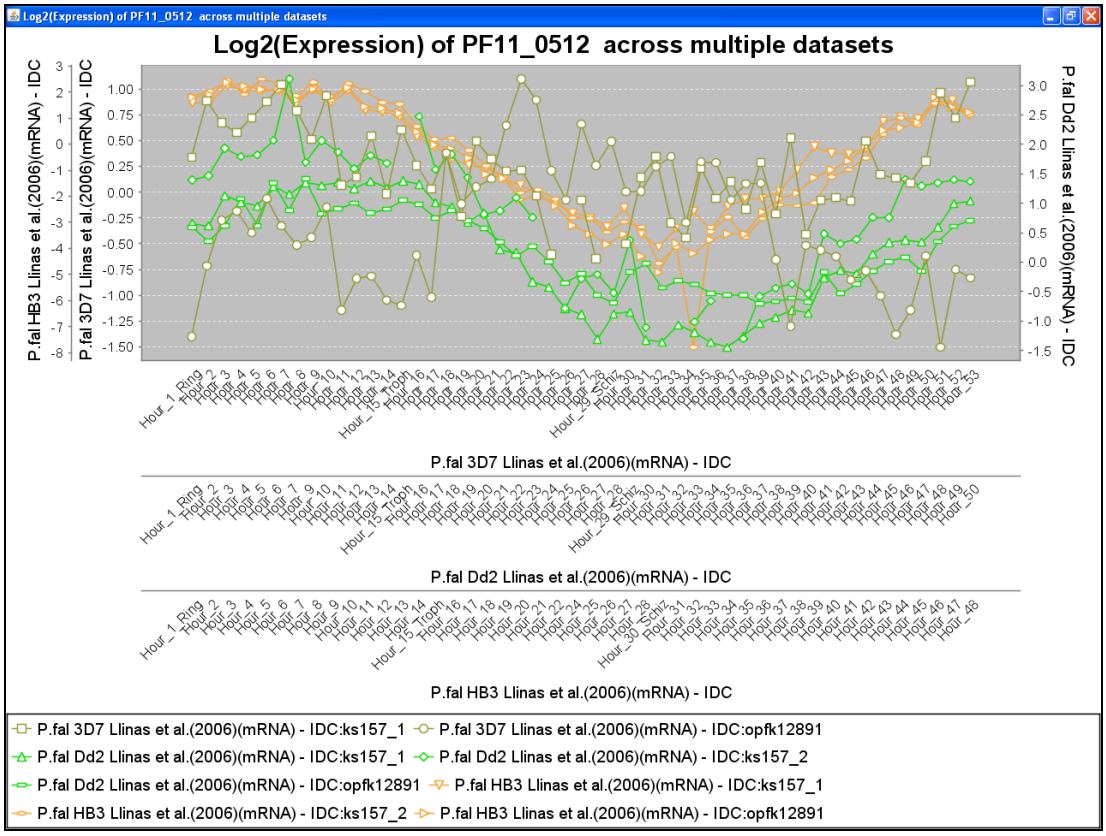


Figure 5.4. Time-series expression profiles for the RESA-2-encoding gene PF11_0512 in the 3D7, Dd2 and HB3 strains. The gene expression profile appears similar between Dd2 and HB3 but varies between these two strains and the 3D7 strain. There is also much variation between different oligos from within the 3D7 gene.

Llinas et al. also reported that gene PFC0110w encoding cytoadherence linked asexual protein (CLAG 3.1) had large changes in its expression profile between strains. Figure 5.5 shows that the expression profile of CLAG3.1 is very similar between strains 3D7 and HB3; however, the gene does not appear to be expressed in Dd2 parasites. Lack of expression of this protein in Dd2 may be related to other

changes in transcription on chromosome 2 that are linked to disruption of the cytoadherent phenotype (see Section 5.1.2).

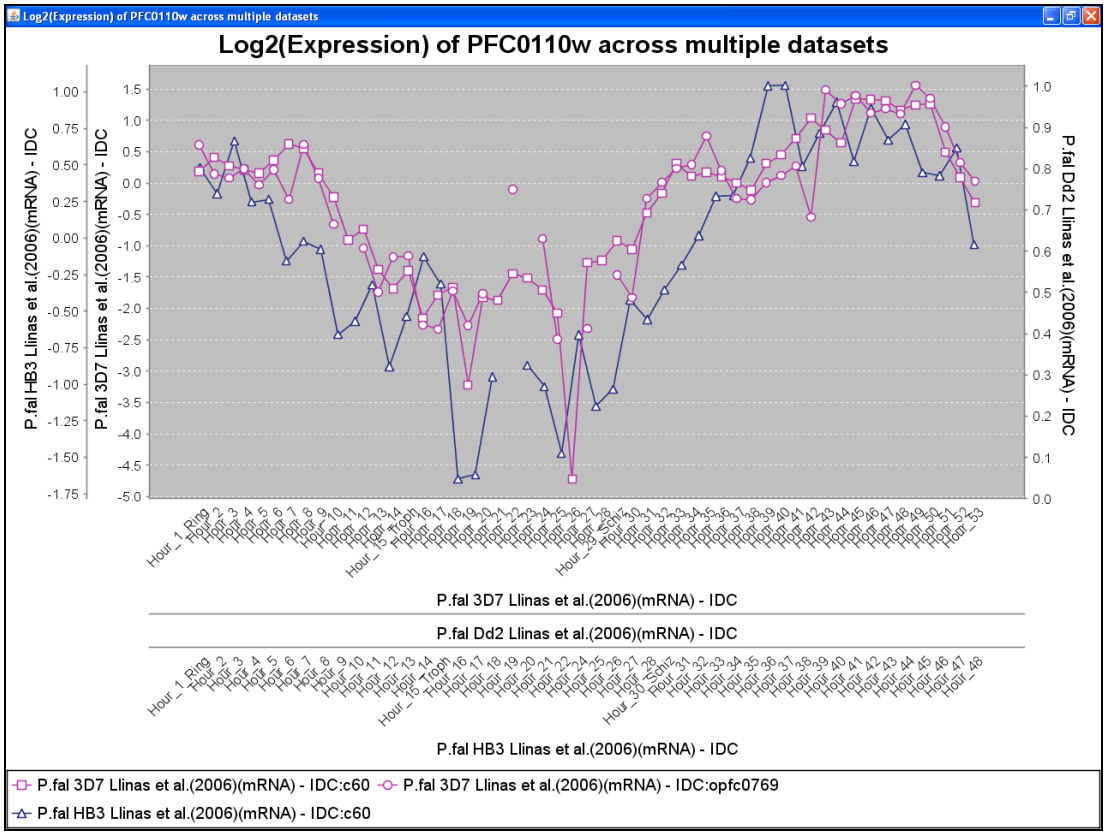


Figure 5.5. Time-series expression profiles for CLAG 3.1-encoding gene PFC0110w in the 3D7 and HB3 strains (no data are available for Dd2). There appears to be little variation in expression profile between the HB3 and 3D7 strains, but expression was not detectable in Dd2 parasites.

Gene PFB0100c, encoding knob associated histidine rich protein (KAHRP), seems to be differentially expressed during the late ring/early trophozoite stages (hours 8-23) in 3D7 parasites (Figure 5.6). Llinas et al. hypothesised that the KAHRP gene may be absent in Dd2 (since the region of chromosome 2 where it is located is known to be silenced or absent in this strain – see Section 5.1.2); however from Figure 5.6 it appears that this gene is present in Dd2 parasites, although data for

one oligo is missing. The authors also stated that this gene is likely to be highly polymorphic in HB3 parasites, although both oligos were detected in this strain at a majority of time-points.

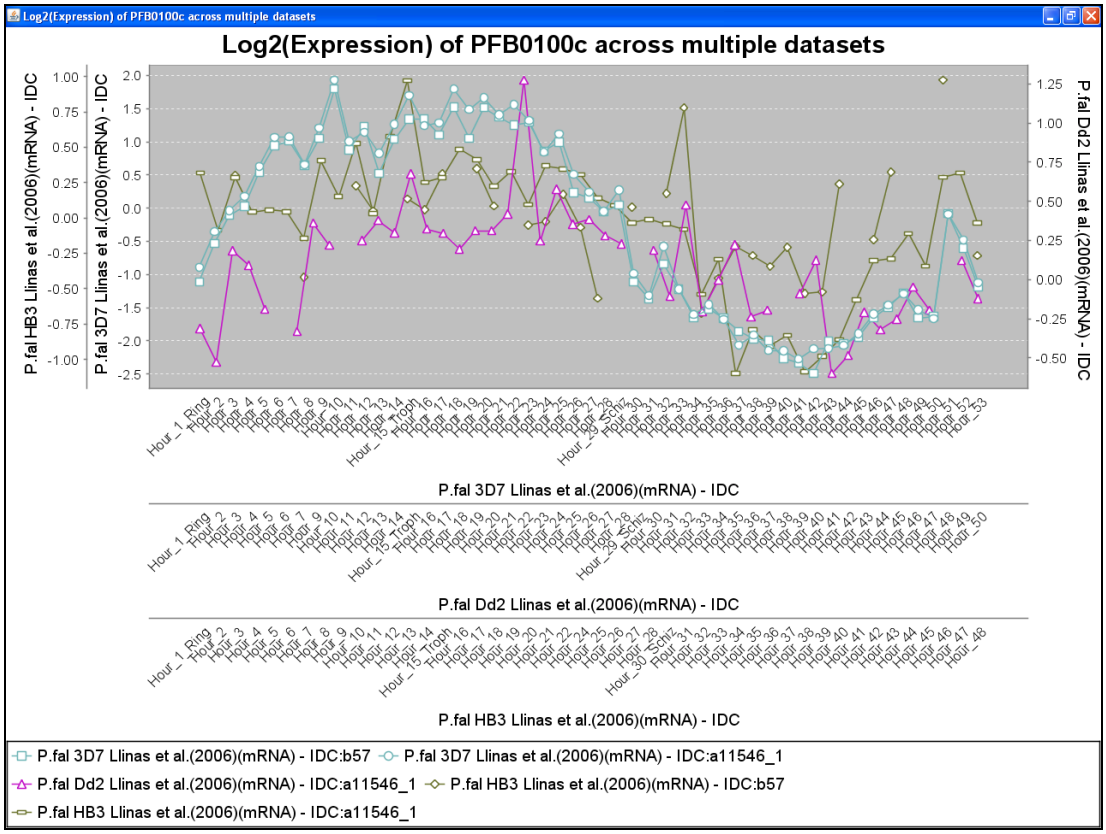


Figure 5.6. Time-series expression profiles for KAHRP-encoding gene PFB0100c in the 3D7, Dd2 and HB3 strains. The gene appears to be differentially expressed during late ring/early trophozoite developmental stages (hours 8-23), which correlates with its known function as a knob surface protein.

A KAHRP neighbour, PFB0095c, which encodes erythrocyte membrane protein 3 (*Pf*EMP3) and is differentially expressed in the IDC of 3D7 strain parasites, recorded no expression signal at all in strains Dd2 and HB3 (Figure 5.7). Llinas et

al. hypothesised that this gene may also be absent in Dd2 and highly polymorphic in HB3, for which Figure 5.7 demonstrates there is good evidence.

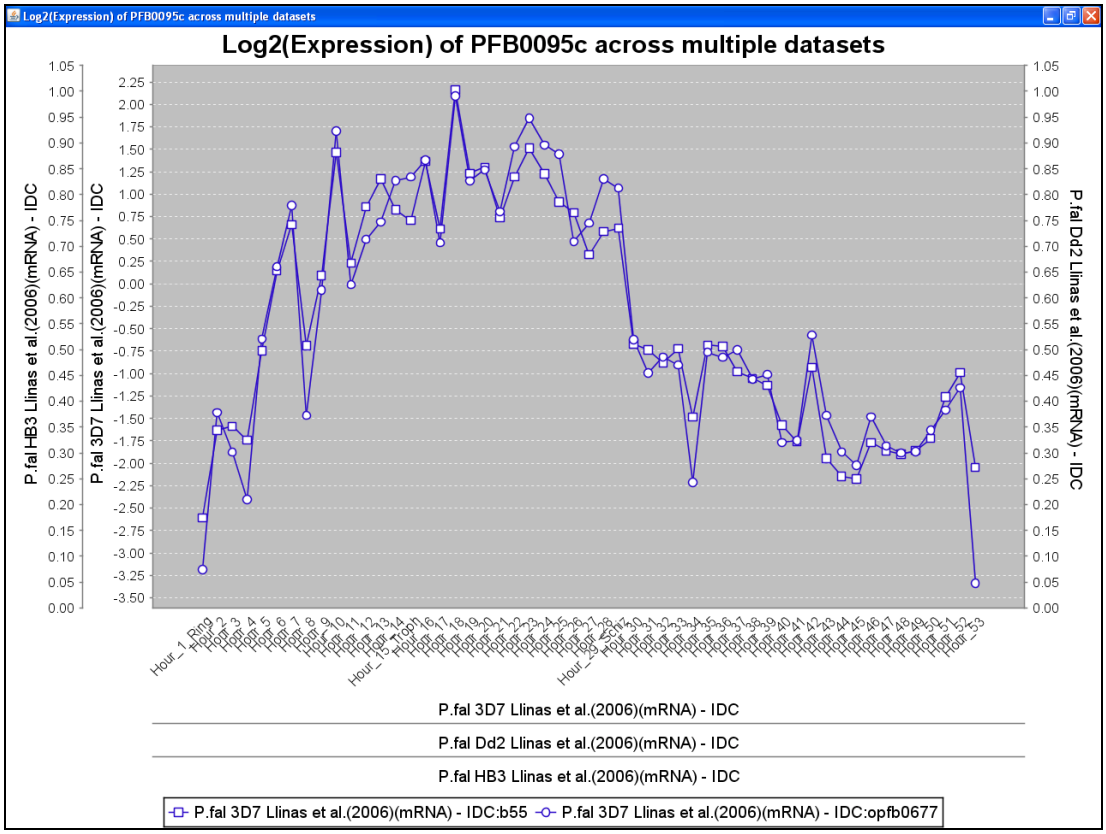


Figure 5.7. Time-series expression profile for the *Pf*EMP3-encoding gene PFB0095c in the 3D7 strain (no signal was detectable in Dd2 and HB3 strain parasites).

5.1.1.1 Using MaGnET to identify genes that are differentially expressed during the ring stage and are enriched for GO terms linked to interaction with host

Llinas et al. describe the results of their analysis of GO term enrichment for genes with the greatest difference in expression profile between HB3 and 3D7. They found significant enrichment of extracellular and plasma-membrane associated genes

involved in defence and immunity, adhesion, host-pathogen interaction and antigenic variation.

The following example is presented to demonstrate that MaGnET can be used to screen for differentially expressed genes between strains. In the 'Query Builder' window of the Expression Data Viewer, genes with expression significantly increasing (greater than 2-fold change) between hour 1 (ring) and hour 15 (trophozoite) of the IDC in HB3 were retrieved and added to Group A. Genes significantly decreasing between the same time-points in the 3D7 strain were added to Group B. The genes in both groups were displayed using the Data Analysis Viewer and genes that were present in both Group A and Group B (in other words genes with opposing directionality of expression change in HB3 and 3D7 parasites) were selected and saved to a file. This file now contained a list of genes that were up-regulated during the ring stage of HB3 parasites but down-regulated during the same stage of 3D7 parasites.

The same process was repeated with the conditions swapped; genes with expression increasing during the ring stage of 3D7 but decreasing during the ring stage of HB3 were selected in the same manner and saved to another file. The two files were then combined and enrichment of GO terms within this subset compared to the entire microarray was analysed using the CLENCH2.0 tool (Shah and Fedoroff 2004).

The analysis showed that within the subset of genes with varying expression profiles during the ring stage of 3D7 and HB3 parasites the GO terms that are significantly over-represented include many terms to do with cell adhesion and host-parasite interactions, such as rosetting, cytoadherence, antigenic variation,

pathogenesis, receptor binding, host cell plasma membrane and infected host cell surface knob (Table 5.1). Other significantly over-represented terms included protein regulatory functions, such as protein kinase activity, proteasome complex and endopeptidase activity. These results compare well with those seen by Llinas et al. – enrichment of terms linked to immune evasion and host interaction, although they did not list specific enriched GO terms and the p-values obtained.

Term	Aspect	P-value
Antigenic variation	Biological process	0.008
Cytoadherence to microvasculature, mediated by parasite protein	Biological process	0
Pathogenesis	Biological process	0
Cyclic nucleotide biosynthetic process	Biological process	0.003
Protein amino acid phosphorylation	Biological process	0.022
Rosetting	Biological process	0.001
ATP binding	Molecular function	0.008
Cell adhesion molecule binding	Molecular function	0
Receptor binding	Molecular function	0.044
Glycosaminoglycan binding	Molecular function	0.001
Threonine endopeptidase activity	Molecular function	0.034
Protein serine/threonine kinase activity	Molecular function	0.03
Receptor activity	Molecular function	0.002
Phosphorus-oxygen lyase activity	Molecular function	0.005
Integral to membrane	Cellular component	0.005
Cytoplasmic part	Cellular component	0.004
Intracellular membrane-bound organelle	Cellular component	0.01
Proteasome core complex (sensu Eukaryota)	Cellular component	0.044
Host cell plasma membrane	Cellular component	0.003
Infected host cell surface knob	Cellular component	0.001

Table 5.1. A list of enriched GO terms in genes with varying expression between HB3 and 3D7 ring stage parasites (hours 1-15 of the IDC). At a confidence level of 95%, a p-value of below 0.05 indicates that a category is significantly enriched in this set compared to all *P. falciparum* proteins. [P-values were generated using CLENCH2.0 (Shah and Fedoroff 2004). Full program output is included on the accompanying CD.]

5.1.2 Putative deleted, polymorphic and silenced regions

Transcriptional differences occurring in certain regions of the genome led Llinas et al. to the conclusion that these regions must either be silenced, deleted or highly polymorphic in some strains. One such example is a 20 kb region on chromosome 4 that contains genes encoding reticulocyte binding protein homolog 4 (*Pf*RH4, PFD1150c), erythrocyte binding antigen-165 (EBA-165, PFD1155w) and SURFIN4.2 (PFD1160w). In 3D7 parasites, *Pf*RH4 and EBA-165 are differentially expressed towards the end of the IDC, along with their neighbouring gene *Pf*RH5 (PFD1145c) (Figures 5.8 and 5.9). The timing of their expression fits with their proposed role in merozoite invasion of erythrocytes; *Pf*RH4 is required for the sialic acid-independent pathway of invasion (Stubbs et al. 2005) and EBA family proteins are required for the sialic acid-dependent mode, although EBA-165 itself is thought to be a likely pseudogene, transcribed but not translated (Triglia et al. 2001). In Dd2 and HB3 parasites *Pf*RH4 and EBA-165 appear to be silenced, but the *Pf*RH5 expression profile remains the same as in 3D7 parasites (Figure 5.10). Therefore, transcriptional regulation at this site varies between strains and may be important for the differences in invasion tactics observed between strains. Although the function of *Pf*RH5 is not known (Cowman and Crabb 2006), it appears not to be involved in the same mechanism of sialic acid-dependent to -independent pathway switching.

Llinas et al noted that SURFIN4.2 also appears to be silenced in Dd2 parasites (Figure 5.10); however, as its 3D7 strain expression profile is not the same as the other genes in this region it would seem unlikely for it to be involved in pathway switching (Figures 5.8 and 5.9). Furthermore, when Llinas et al. made their observations the SURFIN4.2-encoding gene PFD1160w was still annotated as a

hypothetical protein. Despite a 2005 paper describing the characterisation of SURFIN4.2 (Winter et al. 2005) the annotation available from PlasmoDB had apparently not yet been updated. Winter et al. noted that SURFIN4.2 was found co-exported with *Pf*EMP1 and RIFIN proteins to the IE membrane during the IDC, but also that it was positioned in an amorphous cap at the parasite apex during merozoite invasion. The Llinas et al. study reported fairly uniform expression of the SURFIN4.2-encoding gene during the IDC (Figure 5.9), which suggests that it is not differentially expressed during the IDC but may be on constitutively. It is intriguing that this gene should also be silenced in Dd2 parasites; perhaps it does have a role in pathway switching after all.

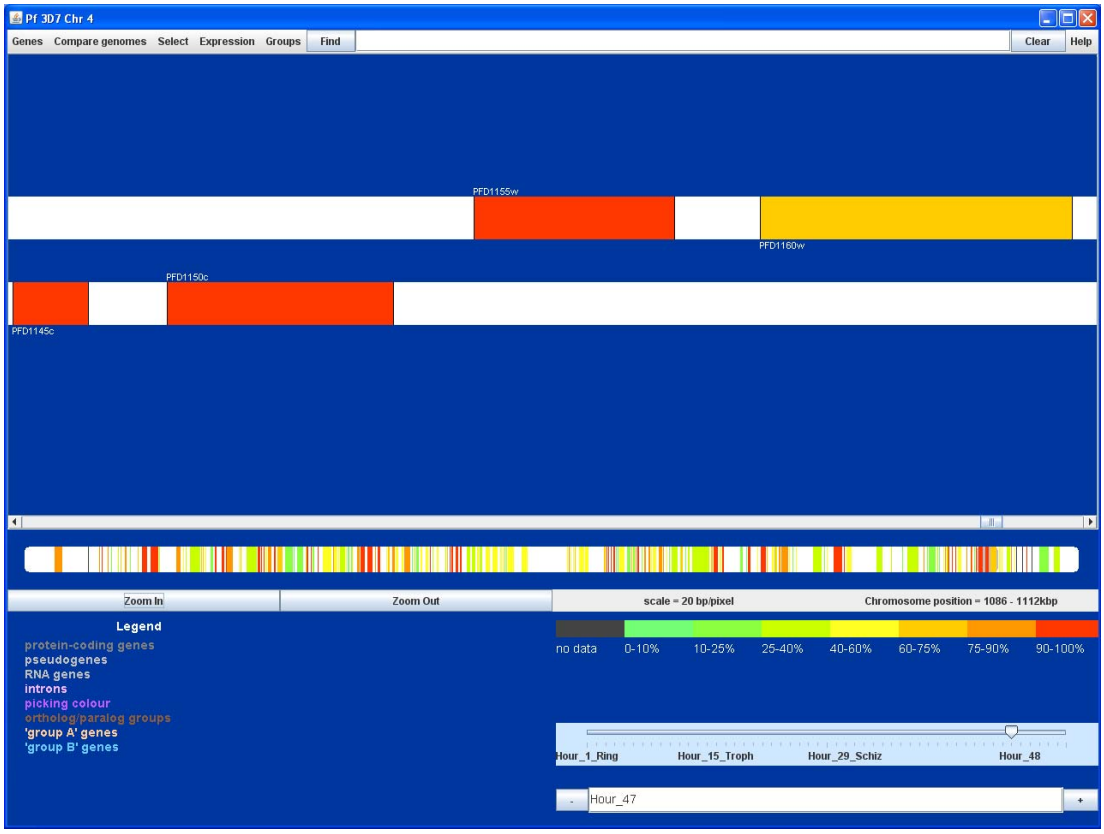


Figure 5.8. A 20 kb region of chromosome 4 containing the genes encoding *Pf*RH5 (PFD1145c), *Pf*RH4 (PFD1150c), EBA-165 (PFD1155w) and SURFIN4.2 (PFD1160w). Expression data at hour

47 of the 3D7 IDC (Llinas et al. 2006) show that the first three genes peak during the latter stages of the cycle (schizogony).

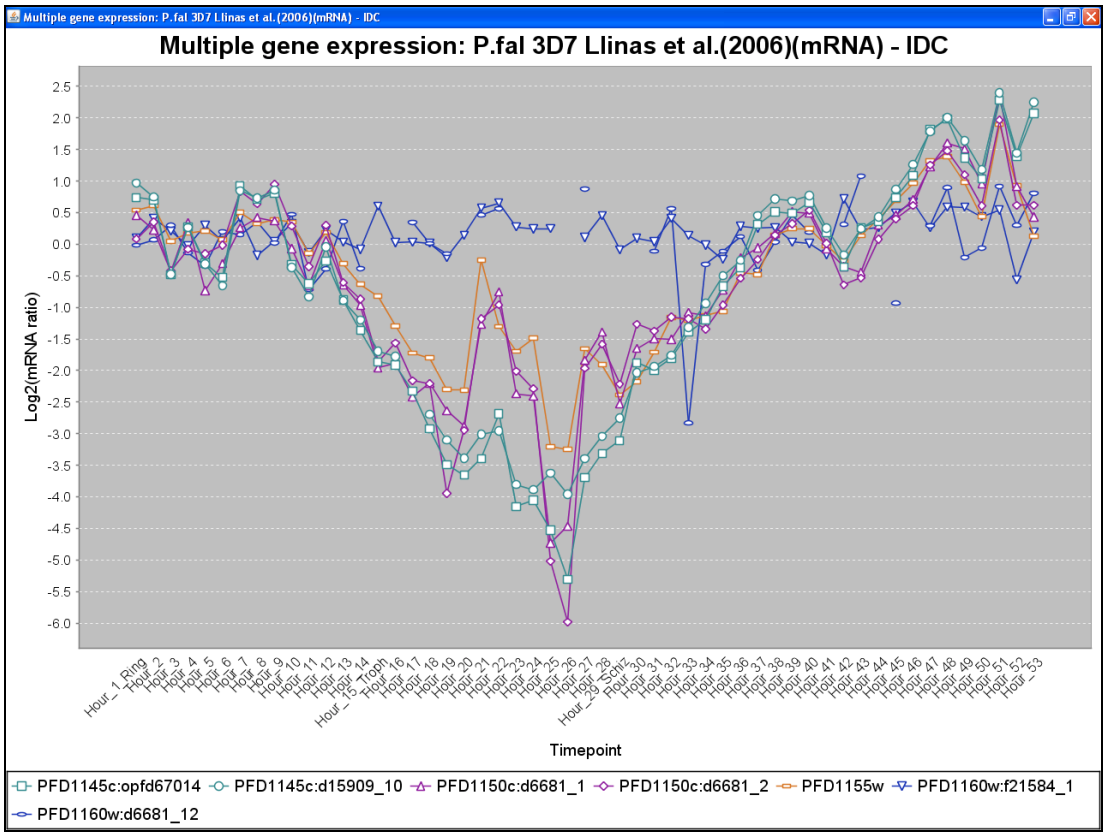


Figure 5.9. Expression profiles of the genes encoding *Pf*RH5 (PFD1145c), *Pf*RH4 (PFD1150c), EBA-165 (PFD1155w) and SURFIN4.2 (PFD1160w) during the 3D7 IDC. *Pf*RH5, *Pf*RH4 and EBA-165 follow a very similar profile, peaking in the late schizont (hours 43-53) (corresponding to the time-frame for generation of merozoites by schizogony).

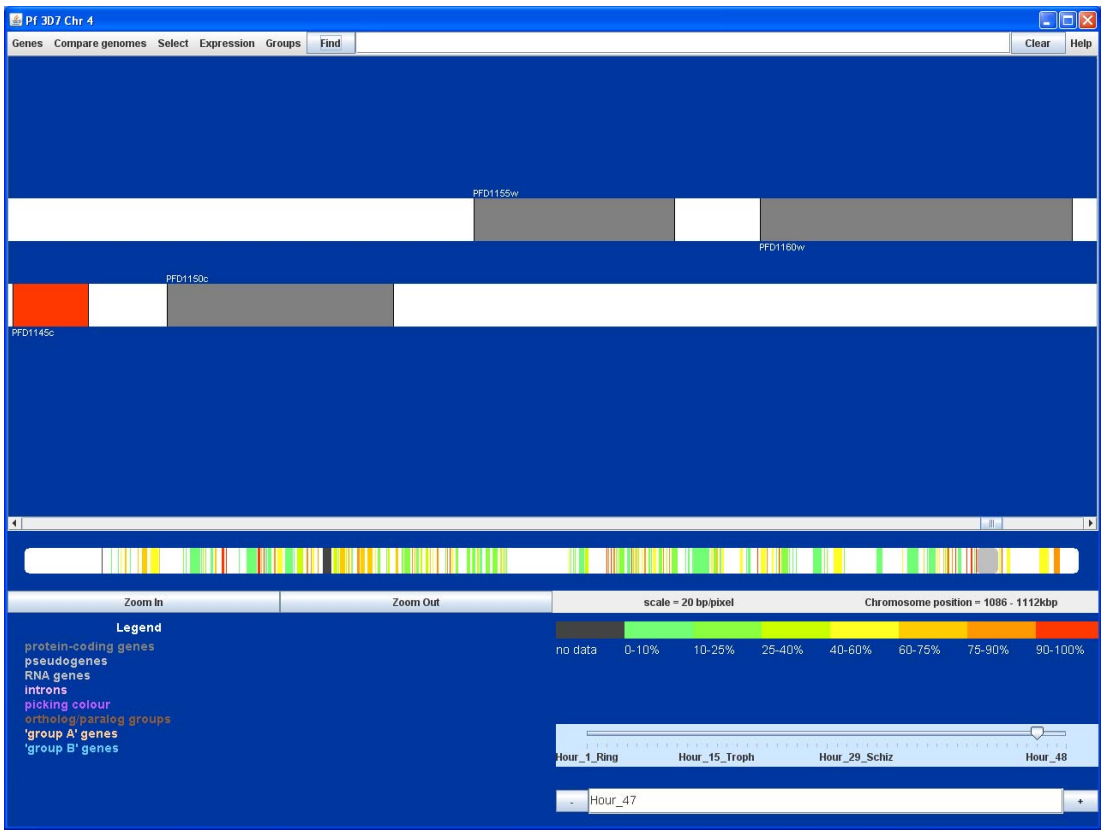


Figure 5.10. A 20 kb region of chromosome 4 containing the genes encoding *Pf*RH5 (PFD1145c), *Pf*RH4 (PFD1150c), EBA-165 (PFD1155w) and SURFIN4.2 (PFD1160w). Expression data displayed are from hour 47 of the IDC in Dd2 strain parasites (Llinas et al. 2006). *Pf*RH5 is the only gene in this region expressed during the Dd2 IDC, leading to the prediction that the other genes in the region are being silenced.

Another region of extreme transcriptional difference between strains encompasses the first 100 kb of chromosome 2, which are known to be involved in cytoadherence (Lanzer et al. 1994). In 3D7 parasites, expression was recorded for many of the genes in this region, including repetitive interspersed family (*rifin*) genes, subtelomeric variable open reading frame family (*stevor*) genes, *var* (encoding *Pf*EMP1), *Pf*EMP3, hypothetical and DNA J domain protein-encoding genes (Figure 5.11). However, in Dd2 parasites expression was recorded for only

one (a DNA J domain protein, PFB0080c) (Figure 5.12). This region is known to be highly polymorphic across *P. falciparum* isolates and deletions in this region are common. Disruption of the knob-associated histidine-rich protein (KAHRP)-encoding gene just upstream is associated with the knobless phenotype of various isolates including Dd2 (Lanzer et al. 1994). Lanzer et al. suggest that a similar phenotype could be caused by either deletions or silencing of transcription occurring within this region of chromosome 2. Either way, it is quite clear from Figures 5.11 and 5.12 that large-scale variations in transcription between strains do occur in the subtelomeric region of chromosome 2.

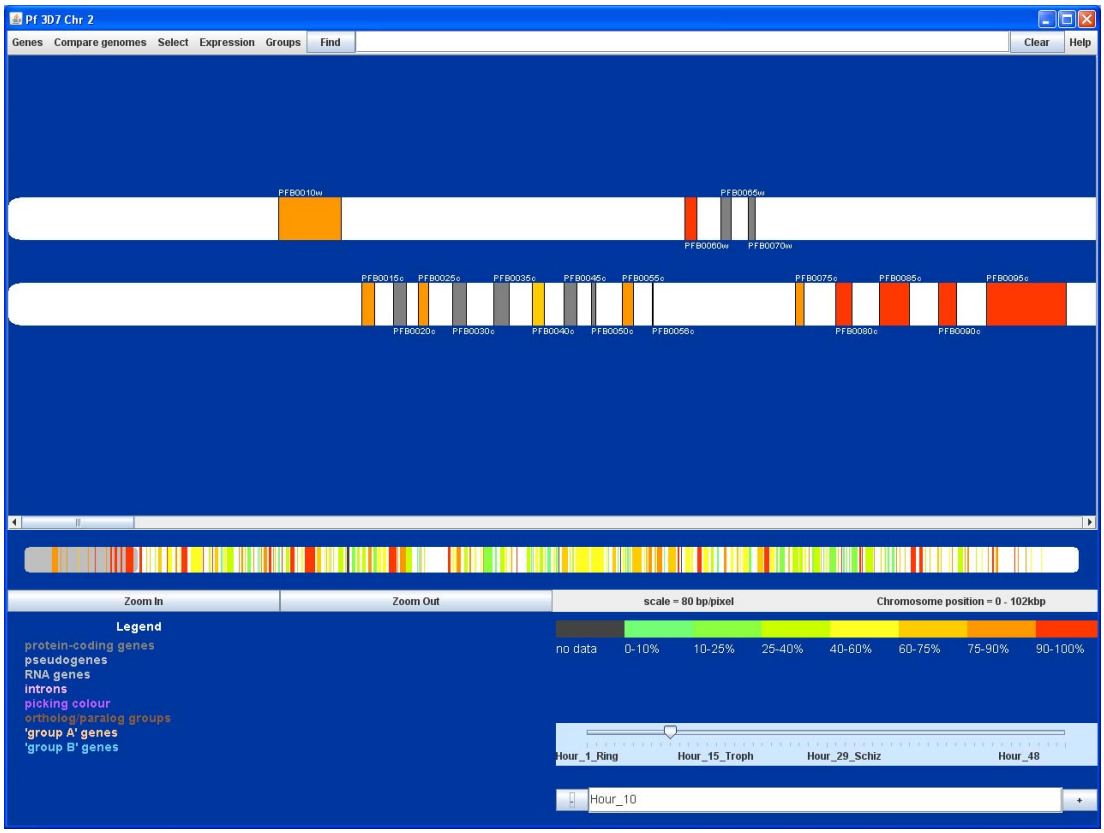


Figure 5.11. The first 100 kb of chromosome 2 displaying expression data from hour 11 of the 3D7 IDC (Llinas et al. 2006).

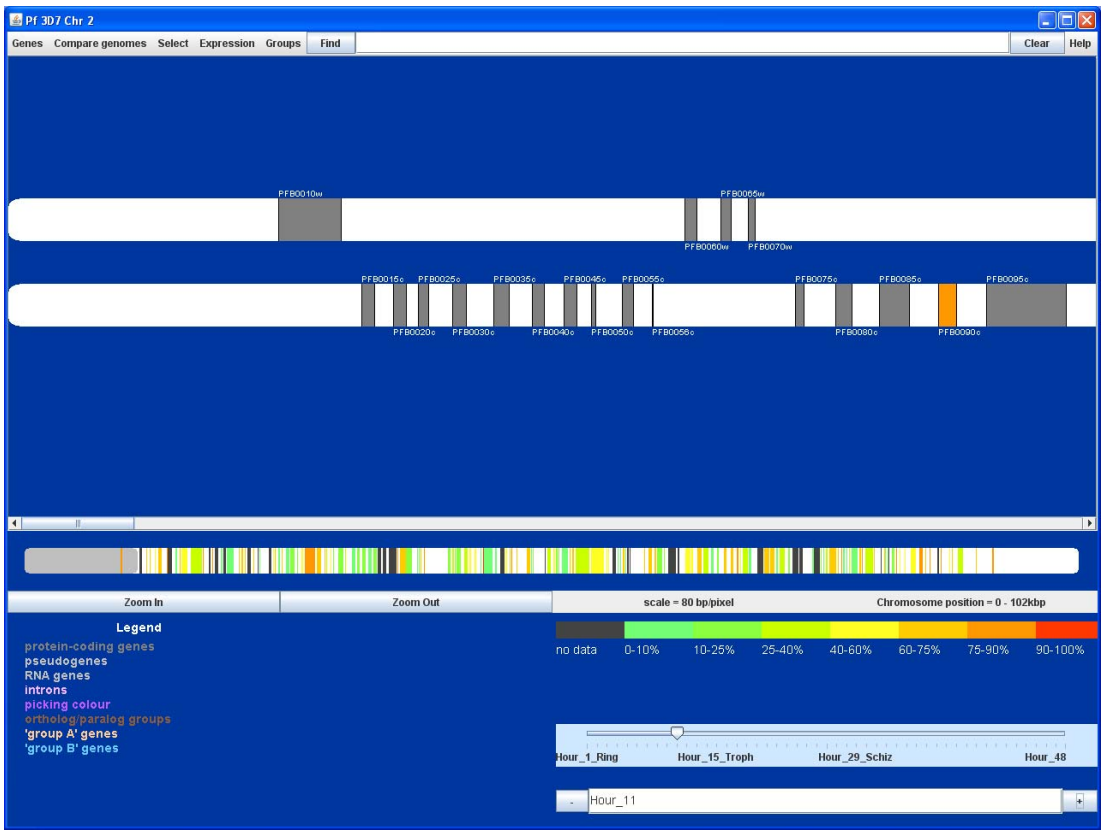


Figure 5.12. The first 100 kb of chromosome 2 displaying expression data from hour 11 of the Dd2 IDC (Llinas et al. 2006).

5.1.3 Immune evasion: *var*, *stevor* and *rifin* genes

The major multigene variant surface antigen families of *P. falciparum* are *var* (encoding *Pf*EMP1), *rifin* and *stevor* genes (Su et al. 1995; Cheng et al. 1998). Llinas et al. reported that their microarray included probes for 50 out of 59 *var* genes, 28 out of 29 *stevors* and 141 out of 154 *rifins*. Figure 5.13 shows the genomic location of the *var*, *stevor* and *rifin* genes and indicates those that have expression data in the 3D7 IDC. Searching for these genes in MaGnET revealed that there are now 73 predicted protein-coding *var* genes, 163 predicted protein-coding *rifins* and 35 predicted protein-coding *stevors*.

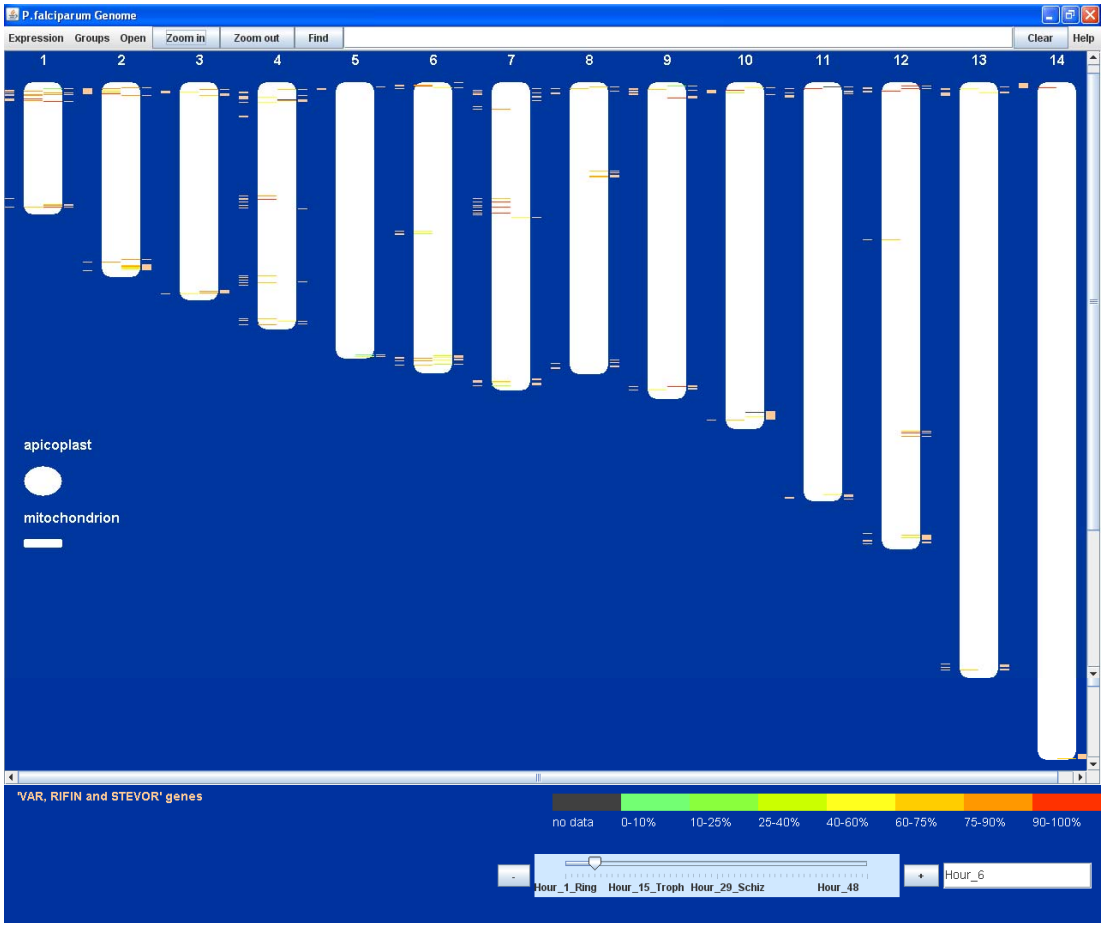


Figure 5.13. Screenshot of the genomic location of *var*, *rifin* and *stevor* genes in the 3D7 strain. Locations of *var*, *rifin* and *stevor* family members are indicated by orange bars next to chromosomes (left side corresponds to reverse strand; right side corresponds to forward strand). 3D7 IDC expression data (Llinas et al. 2006) for any of the genes that meets quality control criteria (data recorded for at least 60% of timepoints) is displayed as a coloured line on either the left (reverse strand) or right (forward strand) half of the chromosome.

Llinas et al. reported that they observed 16 *var* genes to be expressed during the 3D7 IDC. It is not clear what criteria they used to select the *var* genes they considered to be "expressed". There are 39 *var* genes in total with expression data recorded during the 3D7 IDC. The 16 *var* genes they are referring to were probably

considered to be differentially expressed during the IDC. As absolute expression values are not available for the Llinas et al. dataset using MaGnET, instead it was decided to use the expression ratios to estimate the *var* genes that were differentially expressed during the 3D7 IDC. The criterion selected was that a gene's expression ratio must have varied more than 3-fold during the IDC. The Expression Data Viewer Query Builder was used to search for all genes matching these criteria. All the *var* genes matching these criteria (28 genes) were then displayed in the Genome Viewer (Figure 5.14). These 28 *var* genes were considered for the purposes of this investigation to be 'differentially expressed' during the 3D7 IDC. 27 out of 28 of them have maximal expression during the ring and early trophozoite stages (hours 6-22) (Figure 5.14). Interestingly, the 28th gene, PF08_0142, peaks later in the cycle during the schizont stage (hours 29-50). Additionally, the expression timing of the other 27 *var* genes is concurrent with expression of the knob protrusion proteins KAHRP and *Pf*EMP3 (Figure 5.14), correlating with the knob location for *Pf*EMP1.

Comparison of the 28 differentially expressed *var* genes in the Llinas dataset with the 23 highest expressed *var* genes from an earlier study by Le Roch et al. (2003) revealed that less than half the genes overlap between the two sets (Figure 5.15). Therefore, even laboratory strains seem to undergo switching of the subset of *var* genes expressed in different generations. (Note that the Le Roch et al. dataset represents absolute expression levels; therefore direct comparison is not possible.) This result is consistent with that of Llinas et al.

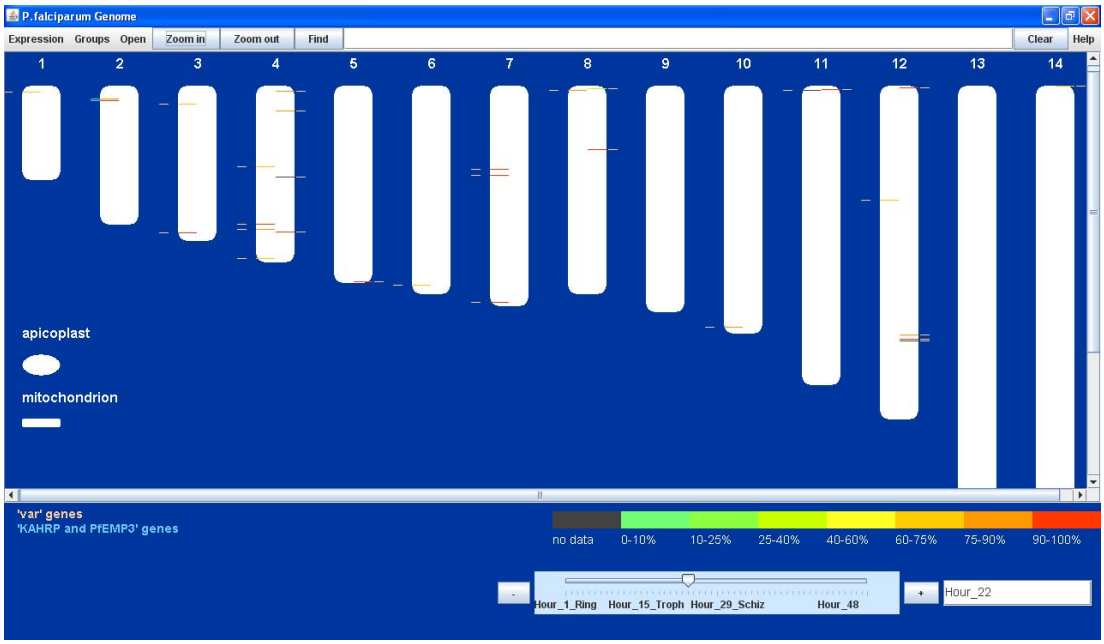


Figure 5.14. Screenshot of the Genome Viewer showing the location of the 28 *var* genes (indicated by orange bars beside chromosomes) that are differentially expressed (undergo greater than 3 fold change in expression) during the 3D7 IDC (Llinas et al. 2006). 27 of the *var* genes peak around the ring/early trophozoite stage (hours 6-22) (maximal expression is indicated by red and orange bars on chromosomes) and the 28th (PF08_0142) peaks in the schizont stage (hours 29-50). Maximum expression of the former 27 *var* genes is concurrent with maximum expression of the KAHRP and *Pf*EMP3 knob protein encoding genes (indicated by blue bars beside chromosome 2).

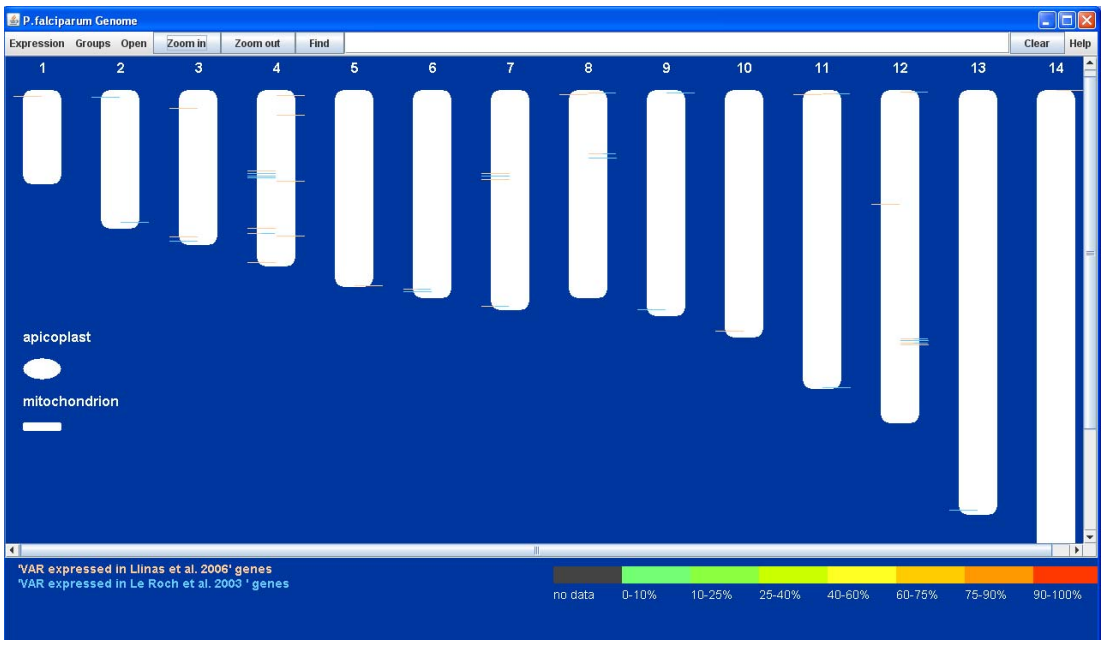


Figure 5.15. Screenshot of the Genome Viewer comparing the location of 28 *var* genes considered to be differentially expressed (expression varied more than 3 fold) in the 3D7 IDC as recorded by Llinas et al. (2006) (orange bars) and the 23 *var* genes with highest expression (absolute expression level higher than 100) as recorded by Le Roch et al. (2003) (blue bars). Bars coloured half orange and half blue represent genes appearing in both sets (totalling 9 genes).

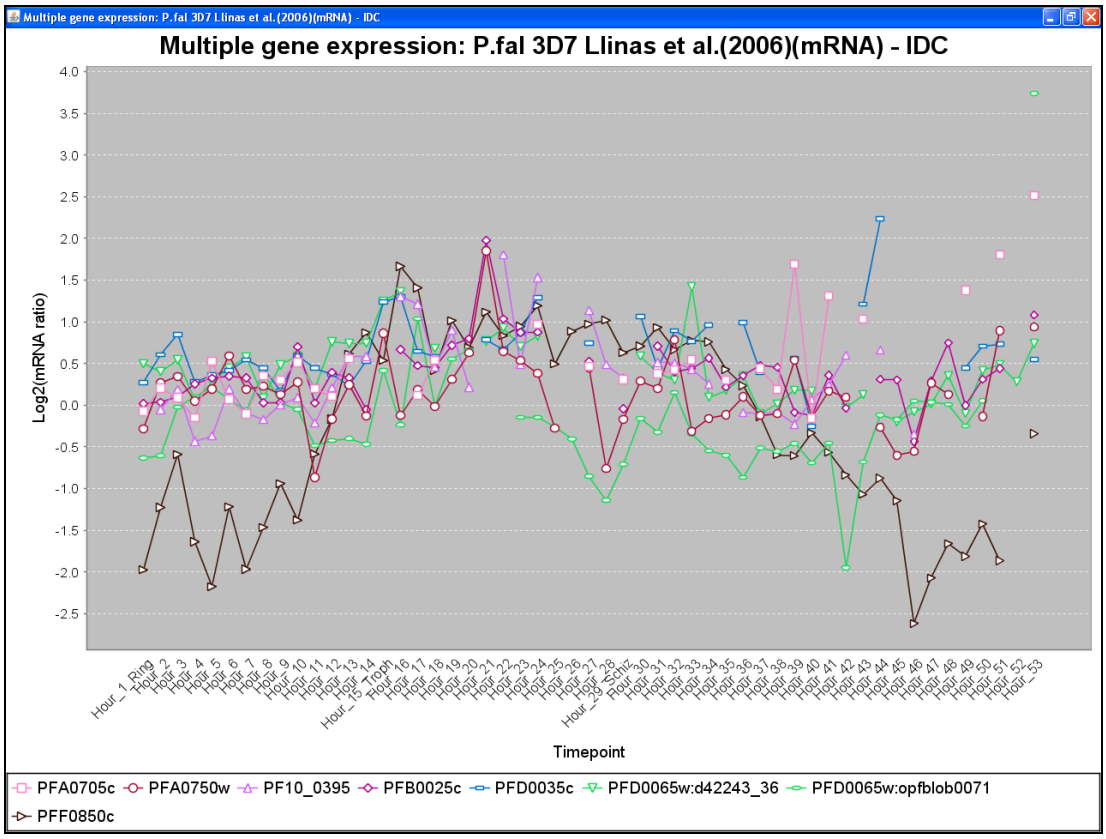


Figure 5.16. Time-series expression profiles of seven differentially expressed *stevor* genes (expression change greater than 3 fold) in the 3D7 IDC (Llinas et al. 2006).

Llinas et al. found little expression of *stevor* genes in the IDC. Figure 5.16 shows that of the 11 *stevor* genes with expression data in the 3D7 IDC, only 7 are estimated to be differentially expressed (using the same criterion as above – that they undergo a 3 fold expression change during the cycle). Their peak expression is generally around the trophozoite stage (hours 14-33), which is consistent with the observed location of their protein products in Maurer's clefts during the schizont stage (hours 29-53) (Kaviratne et al. 2002). The expression of just a small subset of *stevor* genes during the IDC is also further evidence of their proposed differing functional roles in at least three life cycle stages (including gametocytes and sporozoites) (McRobert et al. 2004).

The large multigene *rifin* family is also relatively under-represented in the 3D7 IDC. Out of a total of 49 *rifins* with expression data recorded during the IDC, only 21 of these appear to be differentially expressed (as determined by the criterion that the expression ratio changes at least 3 fold during the IDC). Llinas et al. described a small subset of *rifins* that peak at approximately 15-17 hours post-invasion (early trophozoite stage), which supports their proposed role in adhesion.

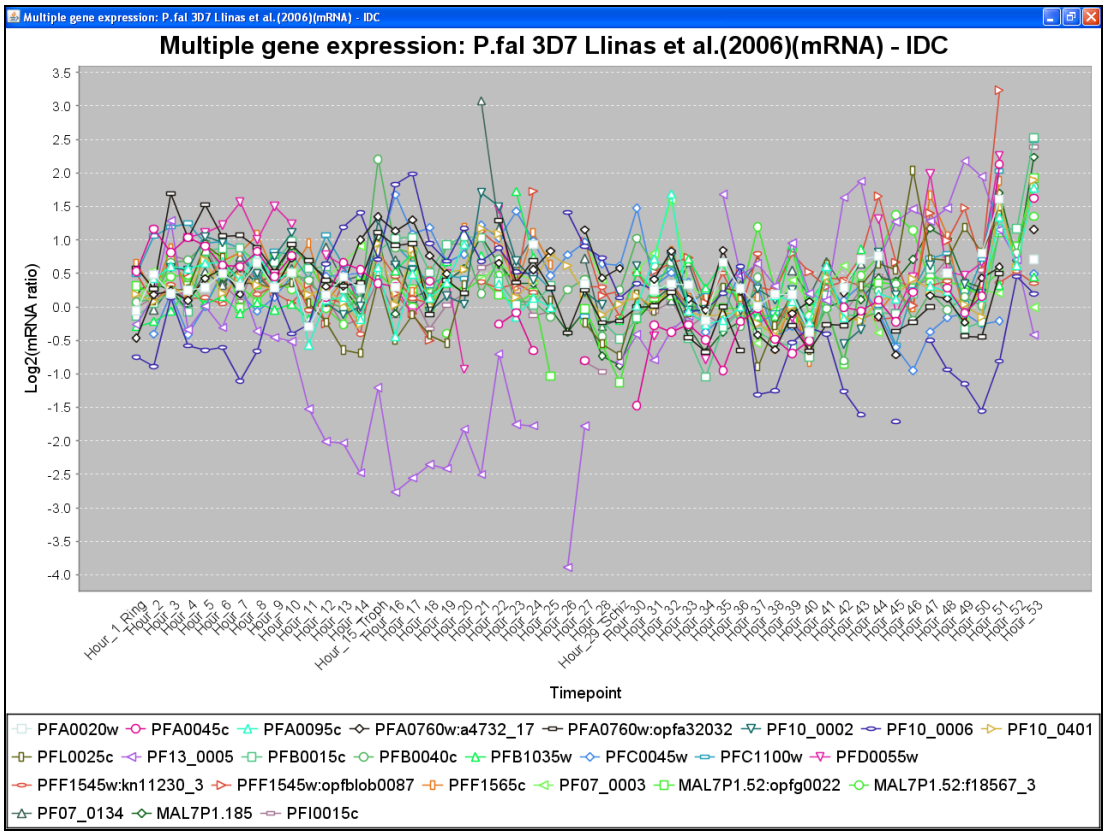


Figure 5.17. Expression profiles of differentially expressed *rifins* (expression change greater than 3 fold) during the 3D7 IDC (Llinas et al. 2006). Several patterns of expression are visible, involving small subsets of genes. One subset of genes peak at around 15-17 hours post-invasion (early trophozoite), other subsets peak at around 20-23 hours (mid-trophozoite) and 30-32 hours (early schizont) and a large peak occurs in the late schizont stage (hours 42-53).

Although this is where the original analysis ended, the analysis of this data using MaGnET was taken a little further. As Figure 5.17 demonstrates, there are several subsets of *rifins* expressed at different points of the IDC. Interestingly, some genes seem to peak towards the end of the cycle during the late schizont stage (hours 42-53). One such gene is PF13_0005, whose expression is down-regulated during the earlier stages of the IDC, and appears to be differentially expressed in mature schizonts. Recently, differing expression and localisation patterns have emerged for two distinct subtypes of *rifin*, termed the A- and B-type *rifins*. A-type RIFINs (the proteins encoded by *rifin* genes) appear to be transported to the surface of infected erythrocytes in asexual stages, whereas B-type RIFINs tend to remain inside the parasite (Petter et al. 2007). RIFINs have also been discovered in the apical region (A-type) and cytoplasm (B-type) of merozoites. Moreover, some family members have only been detected in asexual stages and not merozoites, indicating distinct differential expression patterns (Petter et al. 2007). Therefore, it is not unreasonable to suggest that the PF13_0005 protein may be one of a group of *rifins* differentially expressed in merozoite stages.

5.2 A region of P. falciparum chromosome nine is associated with cytoadherence (Spielmann et al. 2006)

An approximately 55 kb region of the right arm of chromosome nine has been linked to cytoadherence and gametocytogenesis of infected erythrocytes (IEs) [(Spielmann et al. 2006) and references therein]. Spielmann et al. used the available genome sequence of *P. falciparum* 3D7 to pinpoint the exact location of this region

and determine that it contains 13 genes (PFI1715w-PFI1775w) (Figure 5.18). Since none of these genes is *Pf*EMP1, the mediator of IE sequestration, loss of the cytoadherent phenotype caused by loss of function in this region must be due to an indirect effect on *Pf*EMP1 [(Spielmann et al. 2006) and references therein]. In order for *Pf*EMP1 to reach the surface of the IE, the parasite must first set up a protein trafficking network (via the Maurer's cleft) (see Chapter 1), which occurs during the ring stage, along with other modifications of the host cell.

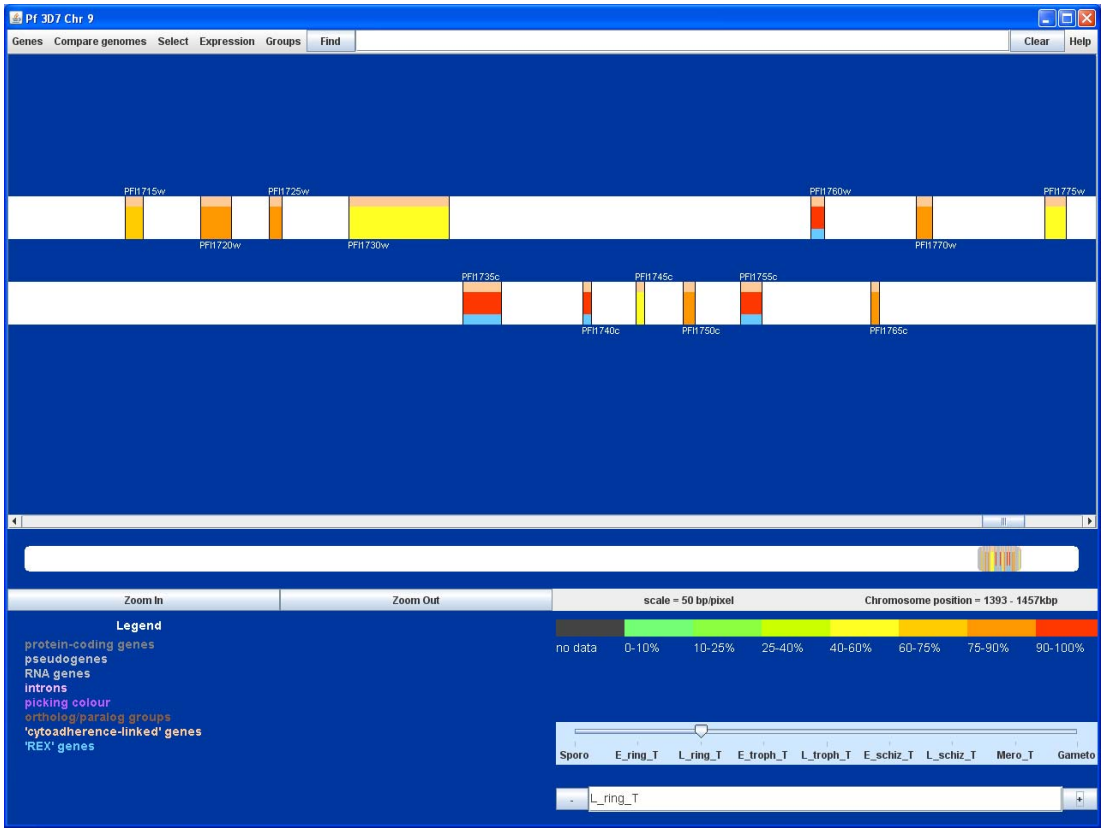


Figure 5.18. Screenshot of a ~55 kb region of the right arm of chromosome 9 linked to cytoadherence and gametocytogenesis. There are 13 genes in this region (marked by orange bars at the top of the gene), four of which have been termed the REX (ring exported) genes due to their ring stage-specific expression and export to the host IE (marked by blue bars at the base of genes). Genes have been coloured according to their recorded mRNA expression level at the late ring stage in an experiment using temperature synchronised cultures of 3D7 parasites) (Le Roch et al. 2003).

5.2.1 A cluster of ring stage-specific genes

Spielmann et al. noted that according to published transcriptome data four of the 13 genes in the described region of chromosome 9 appear to be ring stage-specific (Figure 5.18) (Le Roch et al. 2003). These genes have been termed the ring exported proteins REX1, REX2, REX3 and REX4 (PFI1735c, PFI1740c, PFI1755c and PFI1760w, respectively), since REX1 was previously shown to be exported to the Maurer's cleft (Hawthorne et al. 2004).

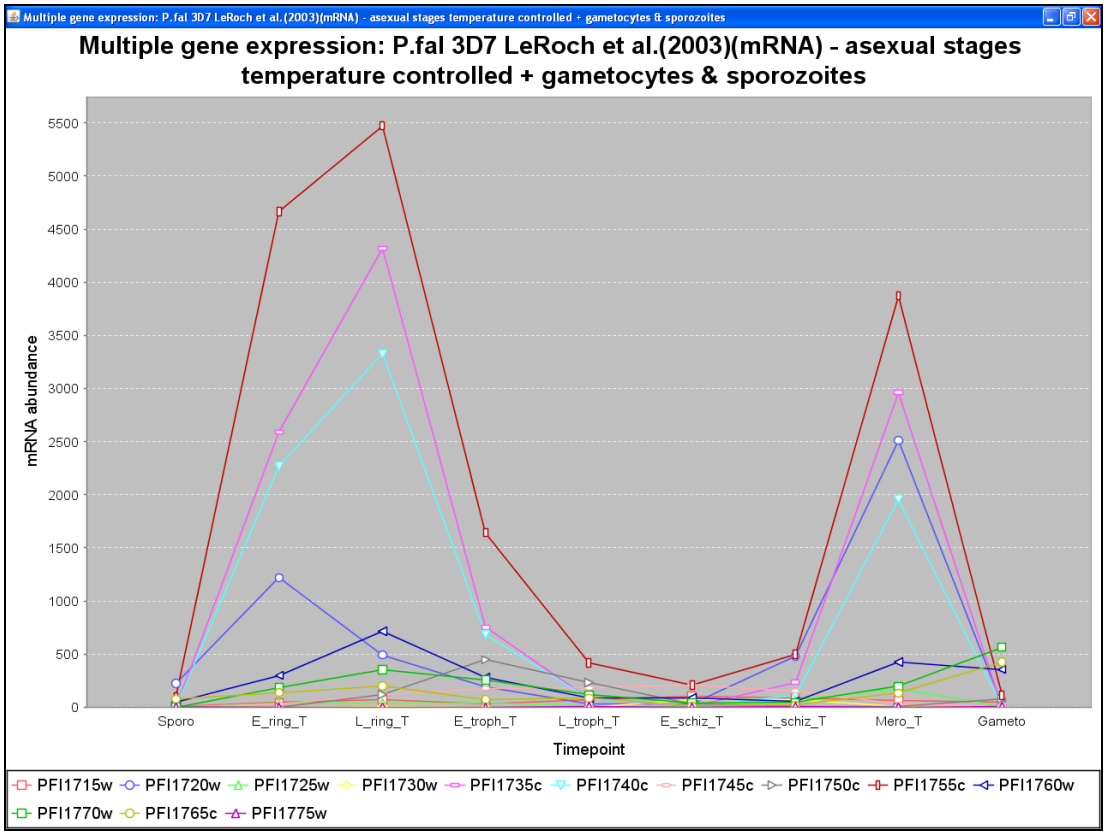


Figure 5.19. Time-series expression profiles of 13 genes in the chromosome 9 cytoadherence locus from 3D7 parasites grown in a temperature synchronised culture (Le Roch et al. 2003). Genes PFI1735c, PFI1740c, PFI1755c and PFI1760w encode the REX proteins REX1, REX2, REX3 and REX4, respectively.

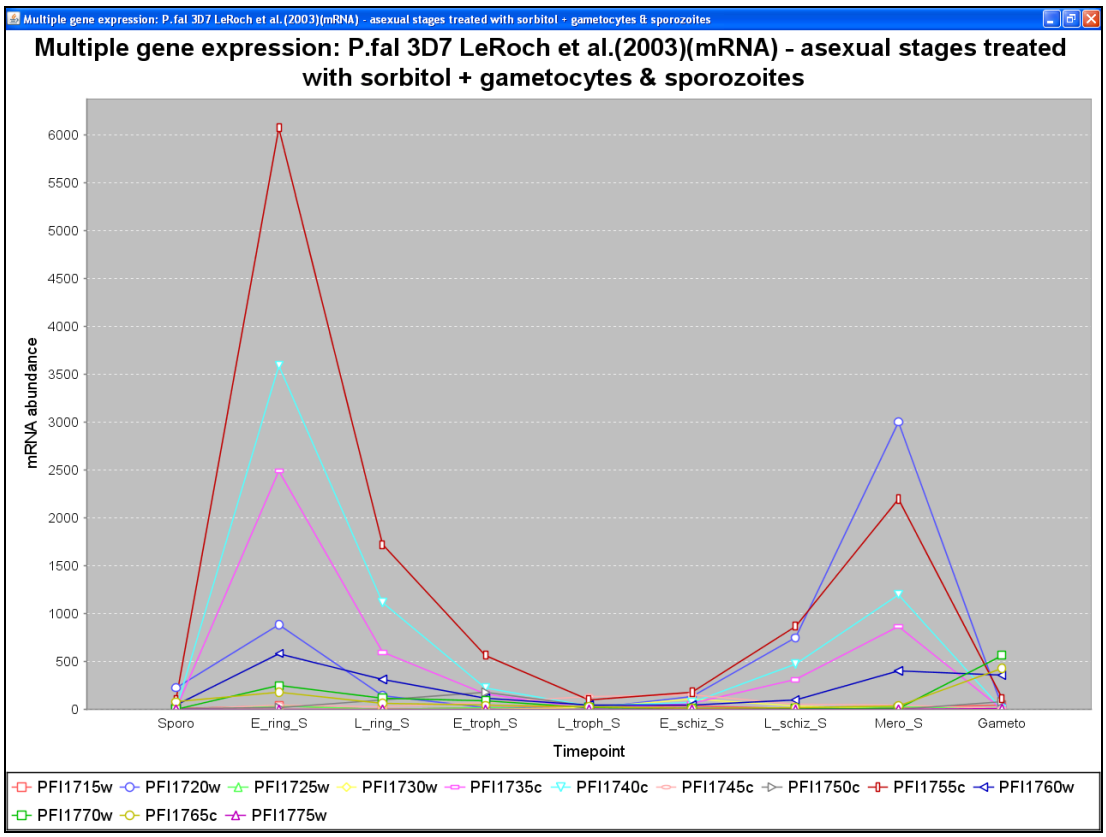


Figure 5.20. Time-series expression profiles of 13 genes in the chromosome 9 cytoadherence locus from 3D7 parasites grown in a sorbitol-treated synchronised culture (Le Roch et al. 2003). Genes PFI1735c, PFI1740c, PFI1755c and PFI1760w encode the REX proteins REX1, REX2, REX3 and REX4, respectively.

Spielmann et al. mentioned that the transcriptome data (Le Roch et al. 2003) showed PFI1765c and PFI1770w were expressed in ring stages (Figure 5.18), but their expression was quite low and increases in gametocytes (Figures 5.19 and 5.20). It should also be noted that based on the above expression profile graphs the gene PFI1720w appears to follow a similar pattern of expression to the REX genes. However, its peak expression actually occurs earlier, during the merozoite stage, whereas peak expression of REX genes occurs during the ring stage. This evidence indicates that PFI1720w should not be clustered with the REX genes. In fact

PFI1720w encodes a protein known as *Pf*gig (*P. falciparum* gene implicated in gametocytogenesis) which is thought to play a role in committing the parasite to gametocytogenesis at the schizont stage of the preceding generation (Gardiner et al. 2005).

5.2.2 REX proteins are encoded by two-exon genes and are unique

Spielmann et al. report that PlasmoDB indicates that each REX gene consists of two exons, a short exon 1 and longer exon 2 (Figure 5.21). They also report that examination of the genes for presence of the PEXEL motif which targets proteins out to the host erythrocyte (Marti et al. 2004) reveals that REX3 and REX4 both contain a PEXEL motif while one is not recognisable in either REX1 or REX2. A PEXEL motif is not always required to direct a protein out to the IE [(Spielmann et al. 2006) and references therein] but a hydrophobic signal sequence does seem to be necessary to target the protein to the parasitophorous vacuole before secretion (Wickham et al. 2001). MaGnET was used to examine whether any of the REX genes have a predicted signal sequence in the current SignalP (Bendtsen et al. 2004) annotation. Interestingly, all four REX genes have a signal anchor sequence predicted at the N-terminus. REX3 and REX4 have signal anchor probabilities of 0.809 and 0.684, respectively, and REX1 and REX2 have signal anchor probabilities of 0.751 and 0.998, respectively (Figure 5.22). Therefore, the protein products of all four REX genes should at the very least be targeted to the parasitophorous vacuole.

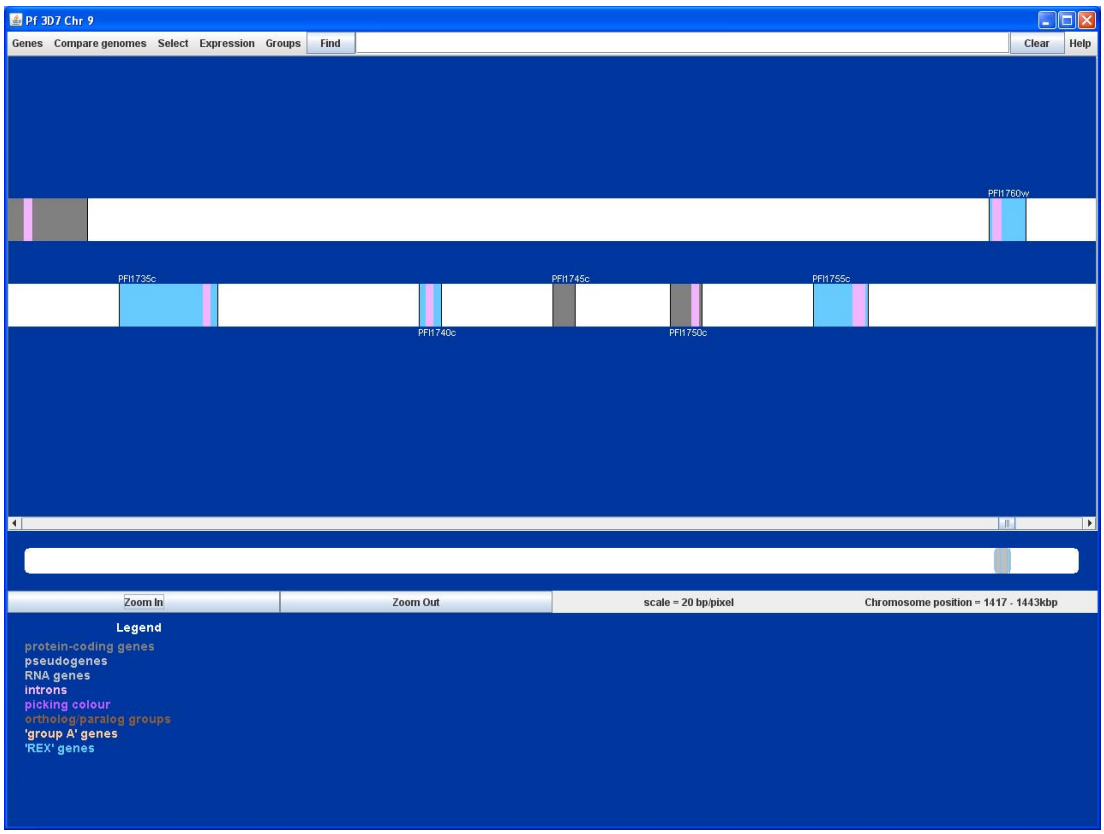


Figure 5.21. Screenshot of a region of chromosome 9 showing the location of introns (pink) in the four REX genes (blue). Each REX gene consists of two exons.

The REX proteins show no similarity to proteins in other organisms. REX3 and REX4 appear to have arisen by gene duplication since they are 31% identical and next to each other on the chromosome (Figure 5.21). The domain shared by REX3 and REX4 is also found in *P. vivax* and *P. knowlesi*. The MaGnET ortholog/paralog group display functionality in the Chromosome Viewer demonstrated that REX 3 and REX4 do indeed have homologs in *P. vivax* and *P. knowlesi*, but not in the rodent species (Figure 5.23).

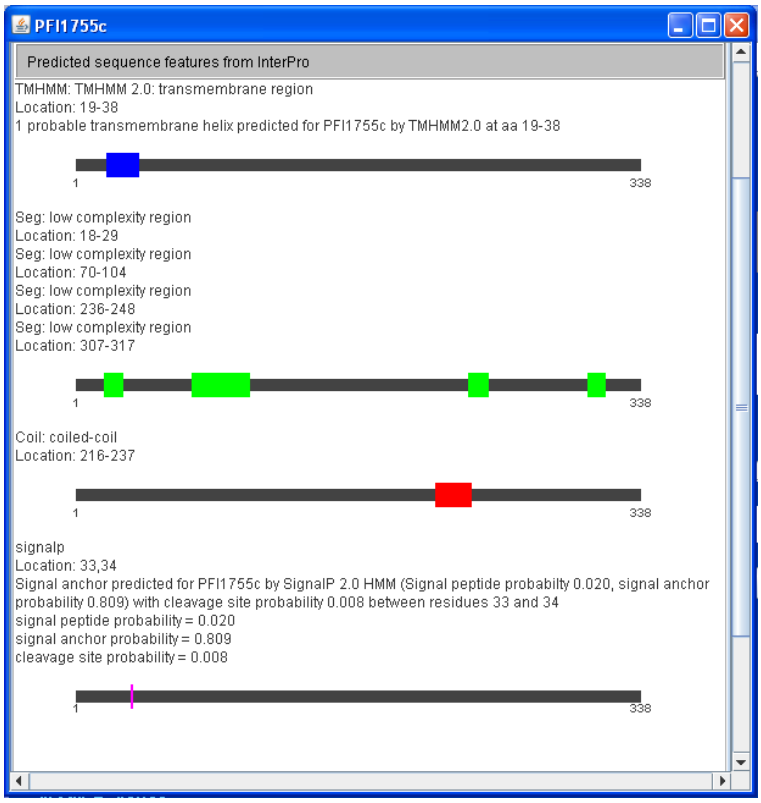


Figure 5.22. Screenshot of the SignalP predicted signal anchor for REX3 (PFI1755c).

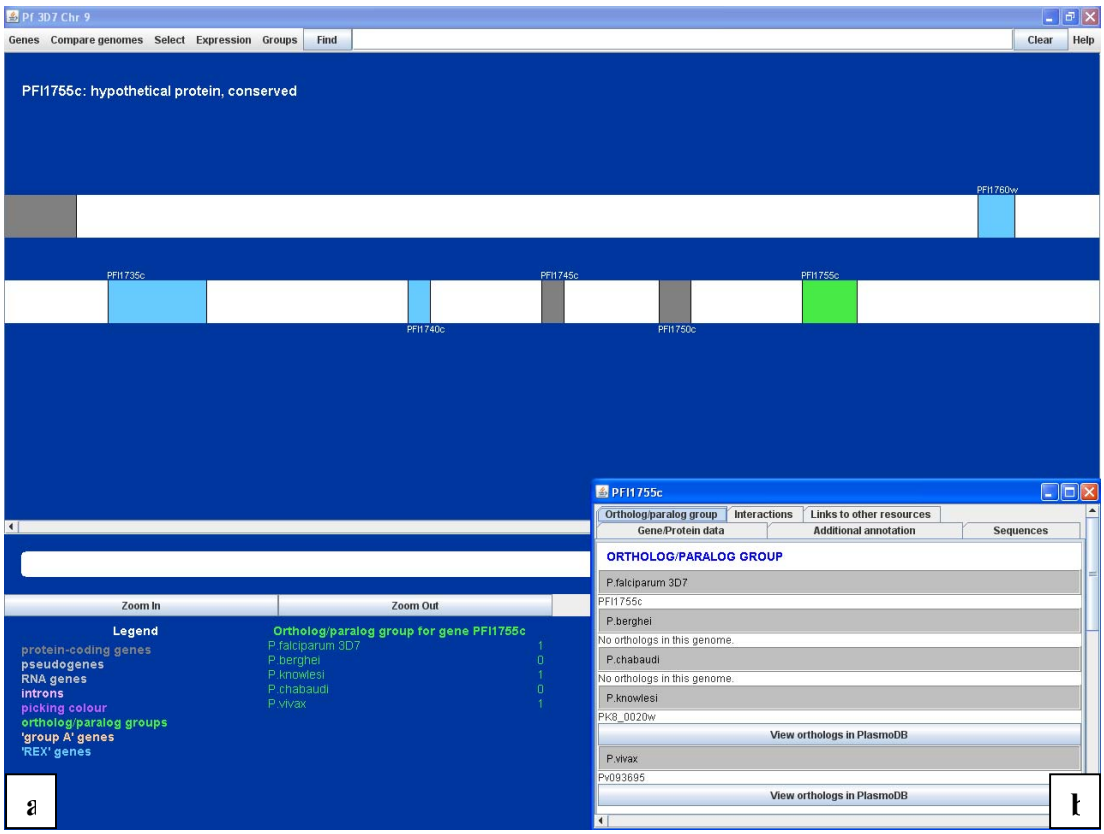


Figure 5.23. Screenshots of the ortholog group for REX3 (PFI1755c): (a) the ortholog/paralog table display in the Chromosome Viewer and (b) the ortholog/paralog group page on the gene fact sheet.

5.2.3 REX1, REX2 and REX3 are exported proteins

REX1 and REX2 are localised to Maurer's clefts and REX3 is found free in the host IE cytoplasm (Hawthorne et al. 2004; Spielmann et al. 2006). However, REX4 protein was not detected and it remains unclear whether this was because it was expressed at very low quantities or whether it is not translated (Spielmann et al. 2006). Visualisation of protein expression data with MaGnET showed that REX1, REX2 and REX3 proteins were all detected in ring stage parasites, whereas REX4 was not detected at all (Figure 5.24).

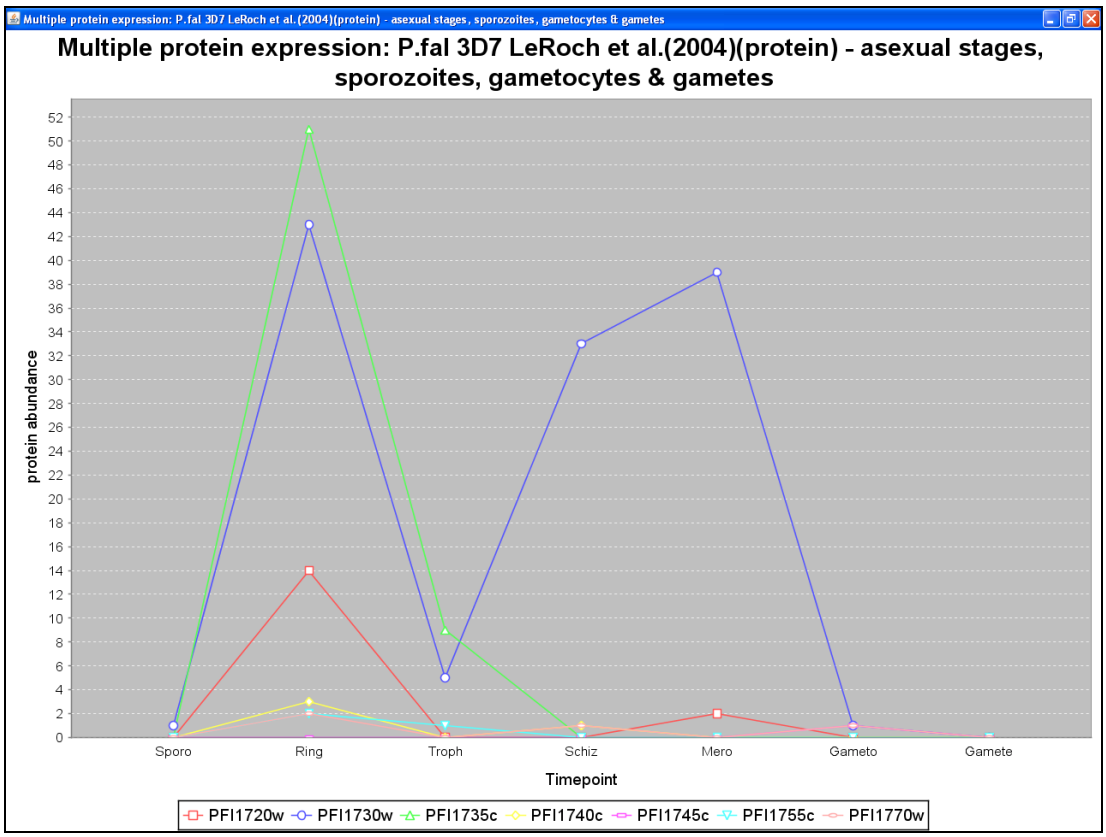


Figure 5.24. Time-series protein expression data for the products of genes encoded by the region on chromosome 9 linked to cytoadherence (Florens et al. 2002; Le Roch et al. 2004). Genes PFI1735c, PFI1740c, PFI1755c and PFI1760w encode REX1, REX2, REX3 and REX4.

5.3 A novel protein kinase family in Apicomplexa (Schneider and Mercereau-Puijalon 2005)

An unusual kinase domain-containing family unique to *Apicomplexa* was first described by Ward et al. in a genome-wide study of *P. falciparum* kinases (Ward et al. 2004). They named the family FIKK after a conserved amino acid motif in the kinase catalytic domain. One member of the family, the trophozoite antigen R45 (PFD1175w), had already been characterised and was found to be associated with the IE cell membrane (Bonnefoy et al. 1992). This protein has been of significant

interest as a possible vaccine candidate. Antibodies to a recombinant protein have been shown to elicit protection against clinical malaria and promote phagocytosis of parasitised erythrocytes (Gysin et al. 1993; Perraut et al. 2003). 20 paralogs of the FIKK family (including R45) were discovered in the *P. falciparum* genome (Ward et al. 2004; Schneider and Mercereau-Puijalon 2005). Schneider and Mercereau-Puijalon described the *P. falciparum* FIKK kinase paralogs in detail in their 2005 paper; the results of their study will be demonstrated in the following sections through the application of various features of MaGnET.

5.3.1 Genomic organisation of FIKK kinase paralogs in *P. falciparum*

The 20 *P. falciparum* paralogs are spread over 11 chromosomes, with a cluster of 7 tandem copies on chromosome 9 and two copies separated by one gene on chromosome 4 (Figure 5.25). 17 of them are located within 150 kb of their telomere.

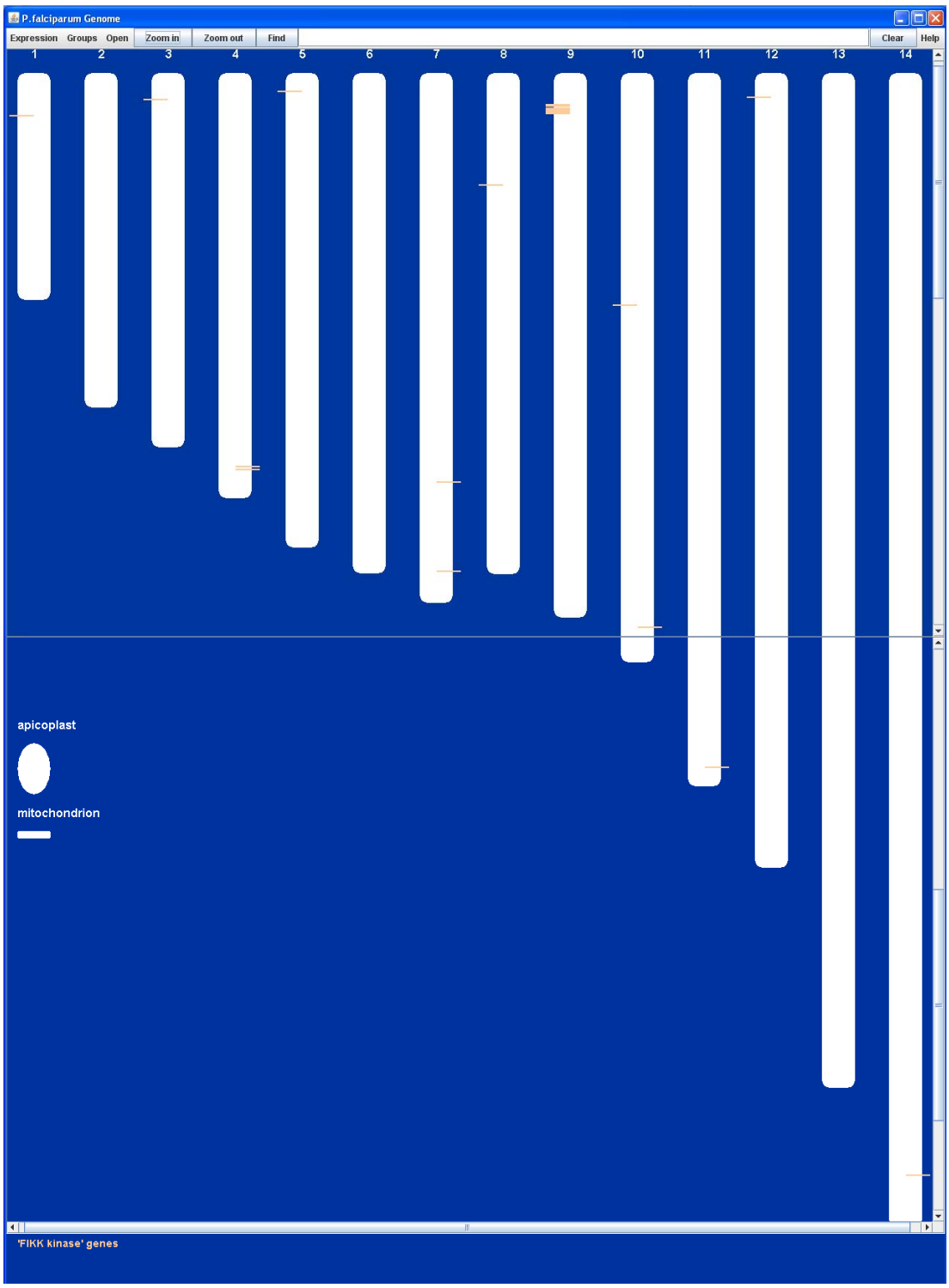


Figure 5.25. The genomic location of 20 FIKK kinase paralogs in *P. falciparum* (orange bars).

5.3.1.1 Exon arrangement

FIKK kinase paralogs have a three-exon gene structure, including a short exon 1, long exon 2 and a short exon 3 (Figure 5.26), except for MAL7P1.175, which lacks exon 1 (or it is fused to exon 2), and MAL8P1.203, which has a short exon 1 and 2 and long exon 3 (Figure 5.27). One of the paralogs appears in the current genome annotation as two separate genes (PF14_0733+4) (Figure 5.28), but sequence alignment shows them to be two parts of a single family member (Ward et al. 2004). The cause of the misprediction presumably came from an internal stop codon. MAL7P1.175 also has an internal stop codon, causing it to be predicted as a pseudogene (Figure 5.27). MAL7P1.175 and PF14_0733+4 may both be pseudogenes, or they may be translated by read-through of the stop codon.

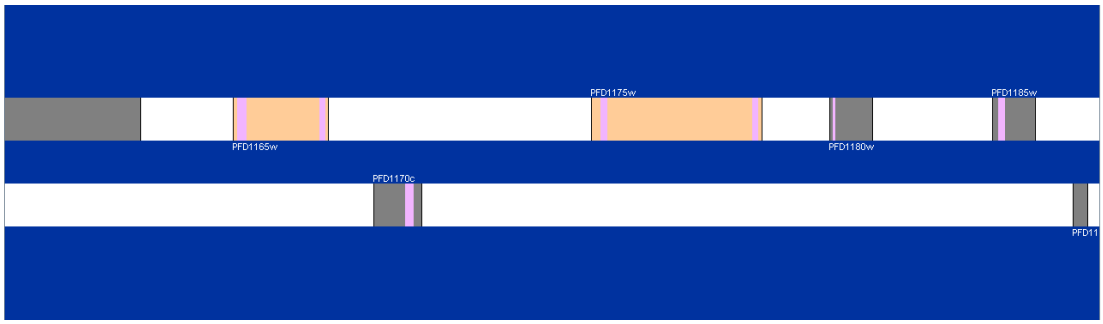


Figure 5.26. Part of the Chromosome Viewer displaying a region of chromosome 4 containing the FIKK kinase paralogs PFD1165w and PFD1175w (encoding R45) (in orange). These genes conform to the typical three-exon gene structure of the family, consisting of short exons 1 and 3 and a long exon 2.

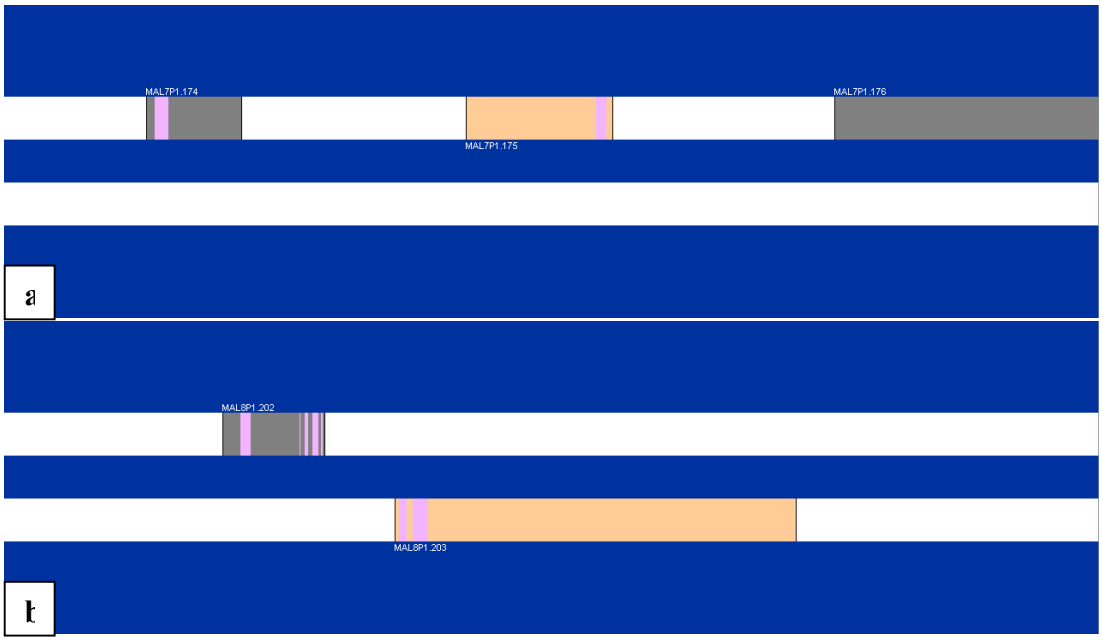


Figure 5.27. Atypical intron/exon arrangements in *P. falciparum* FIKK kinase genes: (a) The pseudogene MAL7P1.175 (in orange) has an atypical gene structure where exon 1 is either missing or fused to the start of exon 2; (b) the gene MAL8P1.203 has a short exon 1 and 2 and a long exon 3, so the short C-terminal exon is either missing or fused to exon 2.

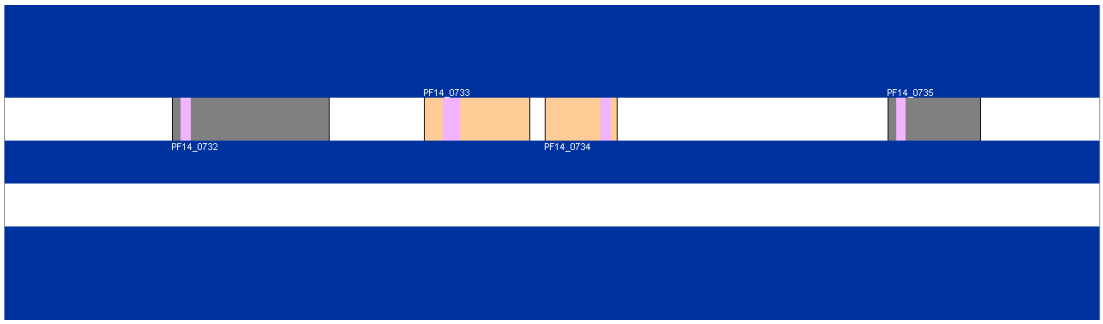


Figure 5.28. Part of the Chromosome Viewer displaying a region of chromosome 14 containing the FIKK kinase family member that was mispredicted as two separate genes (PF14_0733 and PF14_0734) (in orange).

5.3.1.2 All FIKK family members have a conserved C-terminal domain and unique N-terminal region

The FIKK kinase family members share a conserved C-terminal kinase domain, which cannot be assigned to any known kinase family. The kinase domain contains all amino acids necessary for its function apart from the Glycine triad in subdomain I, which is involved in ATP fixation (Ward et al. 2004; Schneider and Mercereau-Puijalon 2005). Nevertheless, the presence of this motif does not appear to be essential for kinase function, which has recently been demonstrated in two of the *P. falciparum* paralogs (Nunes et al. 2007).

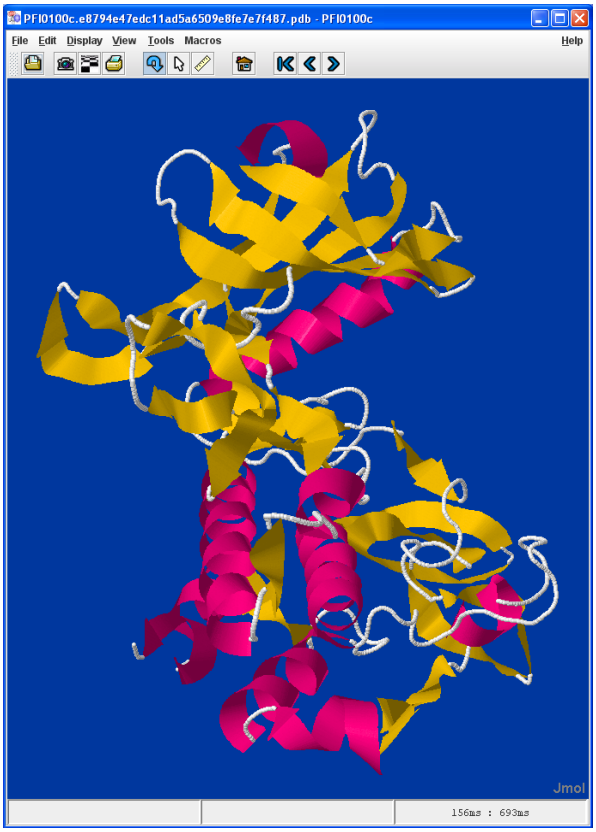


Figure 5.29. A comparative model of the structure of the kinase domain of the protein encoded by gene PFI0100c. The model is displayed as a cartoon representation of its secondary structure.

The functional and structural domain predictions for the FIKK kinase paralogs were explored using MaGnET. 12 out of 20 paralogs have a comparatively modelled structure for a variable-length region of their C-terminal domain. The modelled structures are between approximately 140-340 amino acid residues in length. Figure 5.29 shows a model of the C-terminal domain of PFI0100c, which displays the main structural features of a kinase domain. 19 of the 20 paralogs have at least one hit to an InterPro kinase-like domain (Figure 5.30). The twentieth paralog is the possible pseudogene MAL7P1.175, and there is an overall lack of InterPro annotation for pseudogenes. In comparison, only six of the paralogs have been annotated with GO terms representing kinase functionality.

PFD1175w is distinguished by having a 90 copy hexapeptide repeat inserted between subdomains III and IV of its kinase domain (Figure 5.30). PFI0125c has a stretch of 32 copies of a two amino acid motif in this region, while the other paralogs have a non-conserved stretch of up to 53 non-repetitive amino acids.

The N-terminal region of each paralog is unique and does not contain similarity to any known domain. The N-terminal region is marked by a region of hydrophobic residues corresponding to a likely signal/anchor or transmembrane sequence. Data mining using the MaGnET Data Analysis Viewer revealed signal/anchor sequence or transmembrane domains in 11 out of 20 paralogs (Figure 5.31). Schneider and Mercereau-Puijalon stated that PlasmoDB had listed 14 paralogs with signal or transmembrane sequence annotation at the time of their study. Therefore, during the course of this work, a check was made of the current PlasmoDB annotation and it was found that signal sequences and transmembrane regions were predicted for the same 11 paralogs as in MaGnET. Ultimately,

Schneider and Mercereau-Puijalon manually predicted signal/anchor sequences in all but one paralog. No signal sequence was predicted for PFI0100c, which was later demonstrated to be the only paralog thus far characterised that is not exported beyond the parasite (Nunes et al. 2007).

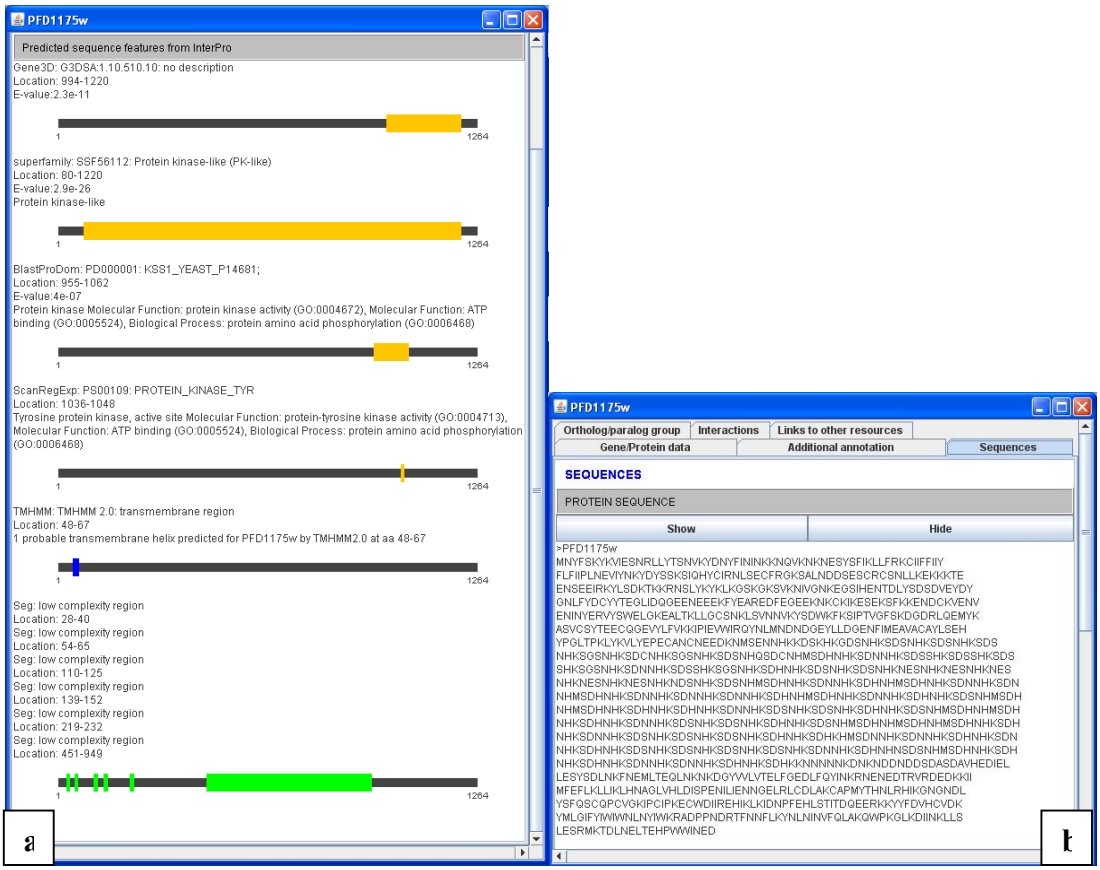


Figure 5.30. Part of the gene fact sheet for R45 (PFD1175w): (a) the InterPro predicted sequence features, showing hits to several kinase-like domains and motifs and a large region of low complexity sequence in the middle of the protein that corresponds to a 90-hexapeptide repeat region (b).

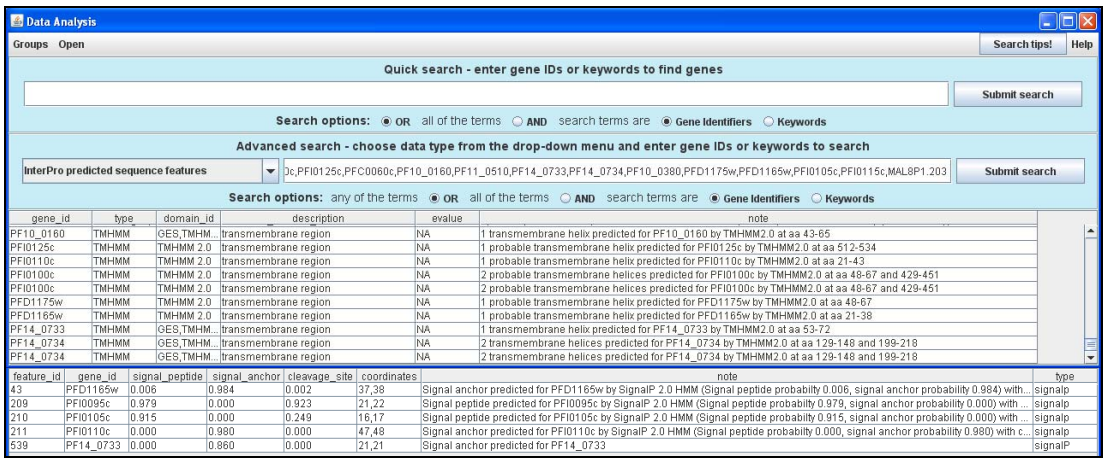


Figure 5.31. Results of searches within the Data Analysis Viewer for the predicted transmembrane domains (top panel) and signal/anchor sequences (bottom panel) for the 20 FIKK kinase paralogs.

5.3.1.3 Subtelomeric FIKK kinase genes are associated with members of other multi-gene families

FIKK kinases are consistently located close to other subtelomeric multi-gene families. 15 out of 17 subtelomeric paralogs are found in close proximity to members of an ortholog group consisting of DNA J domain-containing proteins, which include the ring-infected erythrocyte surface antigens (RESA) (Figure 5.32). Several FIKK kinase paralogs are located next to members of the EBA (Figure 5.33) and fatty acid CoA synthase (Figure 5.34) families. In addition, several multi-gene families coding for hypothetical membrane proteins are found in close proximity to many of the FIKK kinase paralogs (Figure 5.35).

Furthermore, there are examples of specific higher-order arrangements of genes; for instance, RESA, EBA and FIKK kinase genes are often found arranged in tandem (Figure 5.36).

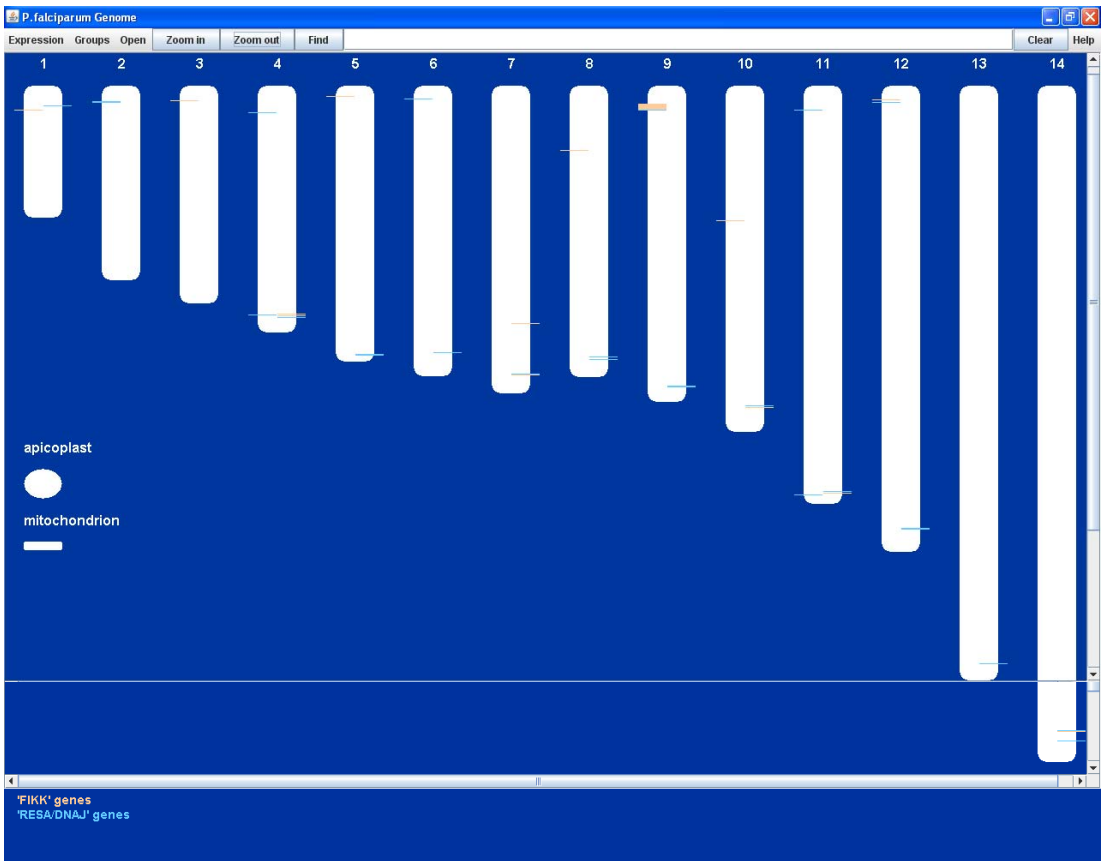


Figure 5.32. Many of the subtelomeric FIKK kinase paralogs (orange bars) are located close to genes in a large multi-gene family coding for DNA J domain-containing proteins (including the RESA proteins) (blue bars).

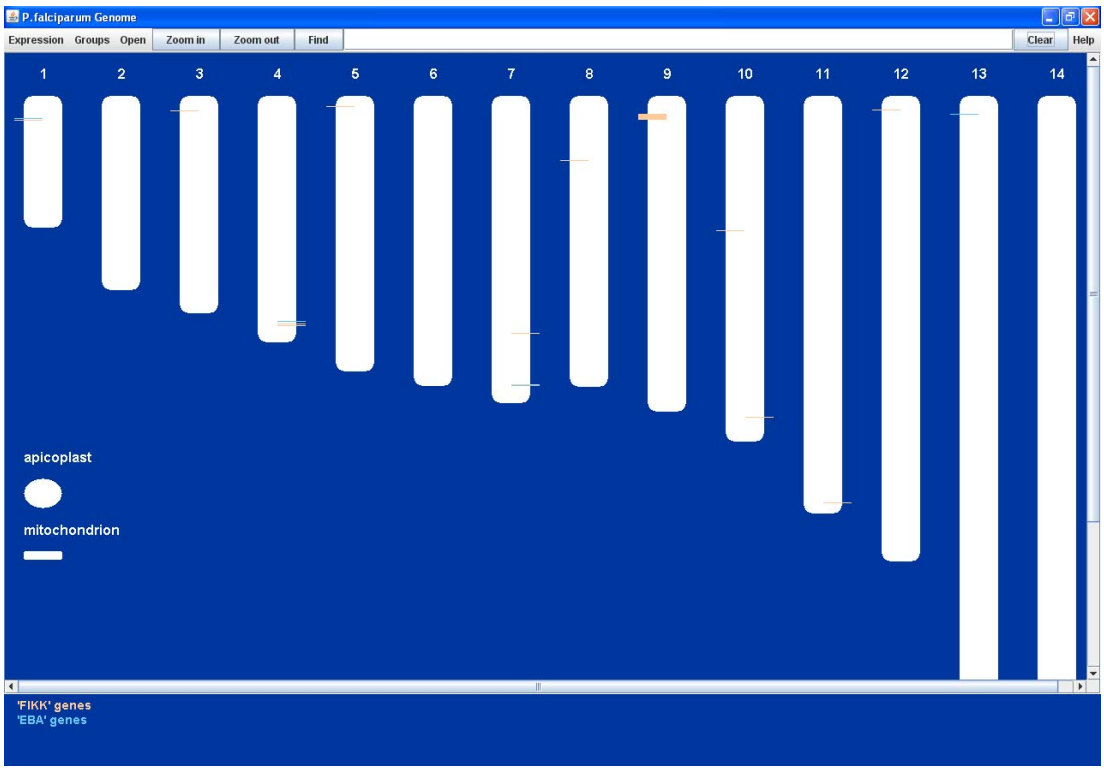


Figure 5.33. Several of the subtelomeric FIKK kinase paralogs (orange bars) are located next to EBA family genes (blue bars).

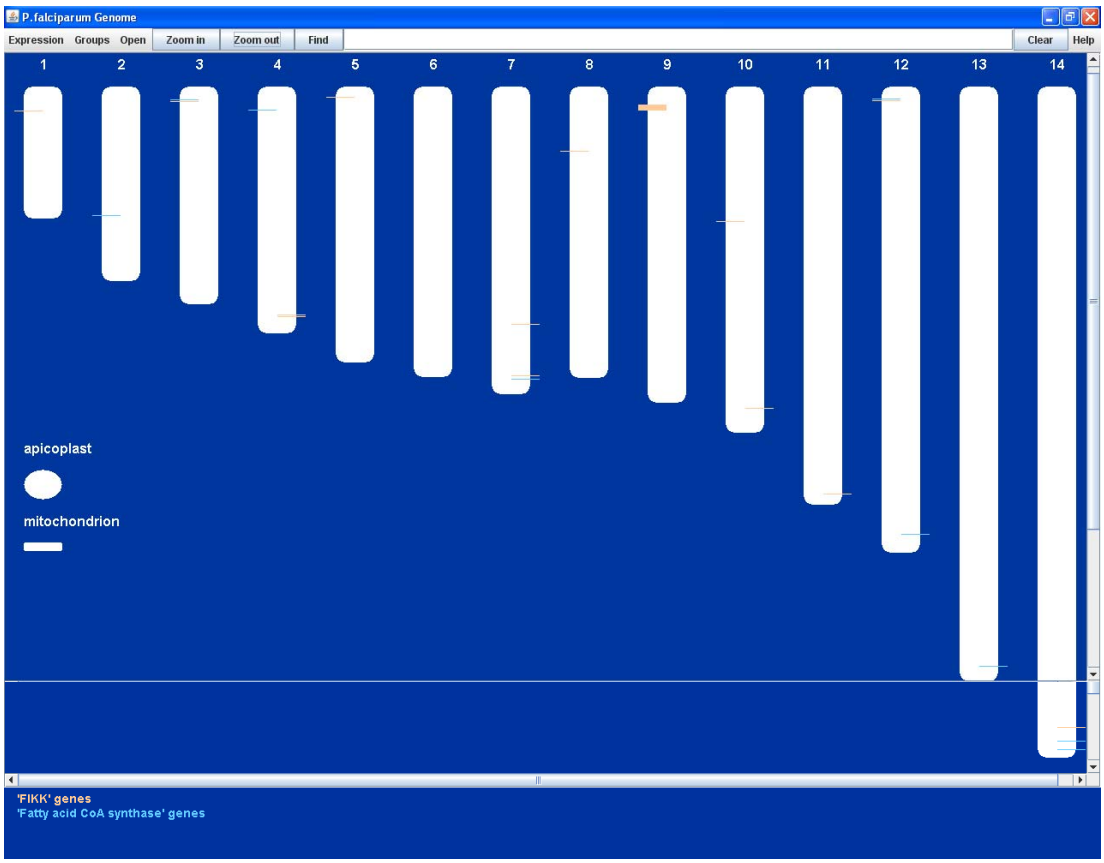


Figure 5.34. Several of the subtelomeric FIKK kinase paralogs (orange bars) are located next to fatty acid CoA synthase genes (blue bars).

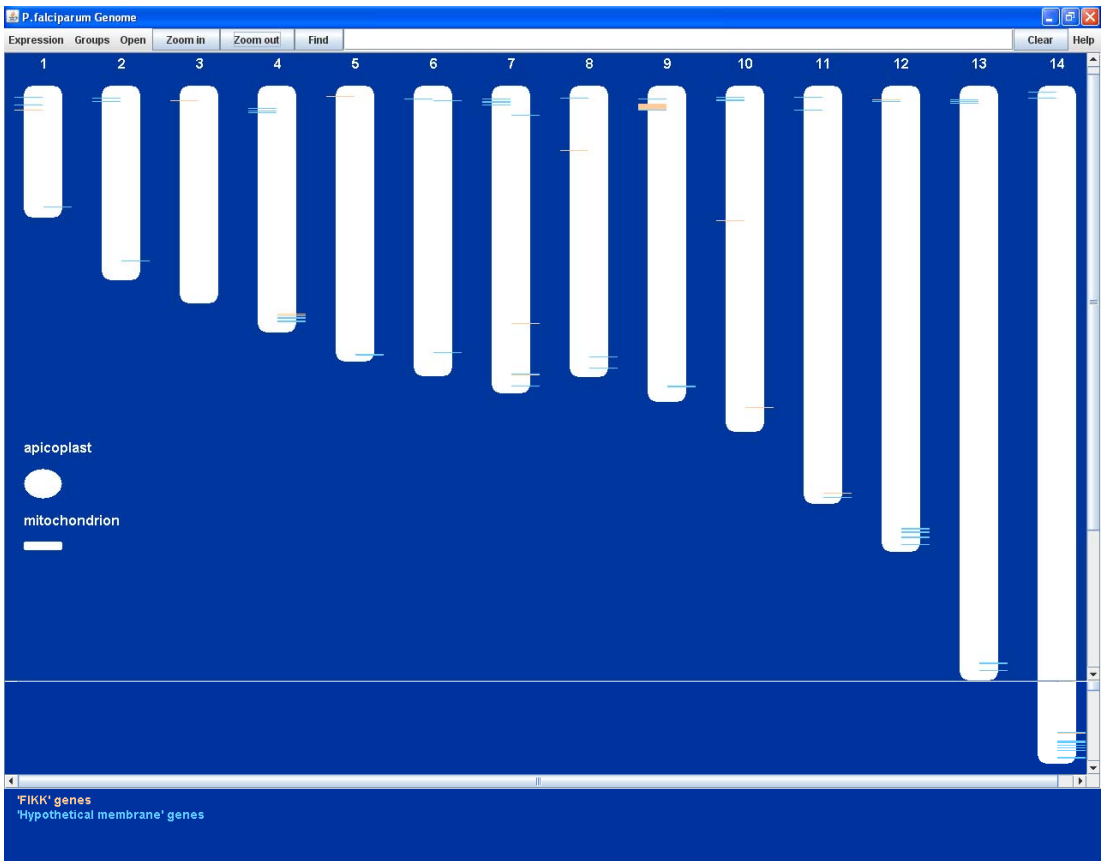


Figure 5.35. Many of the subtelomeric FIKK kinase paralogs (orange bars) are located close to members of several gene families coding for hypothetical membrane proteins (blue bars).

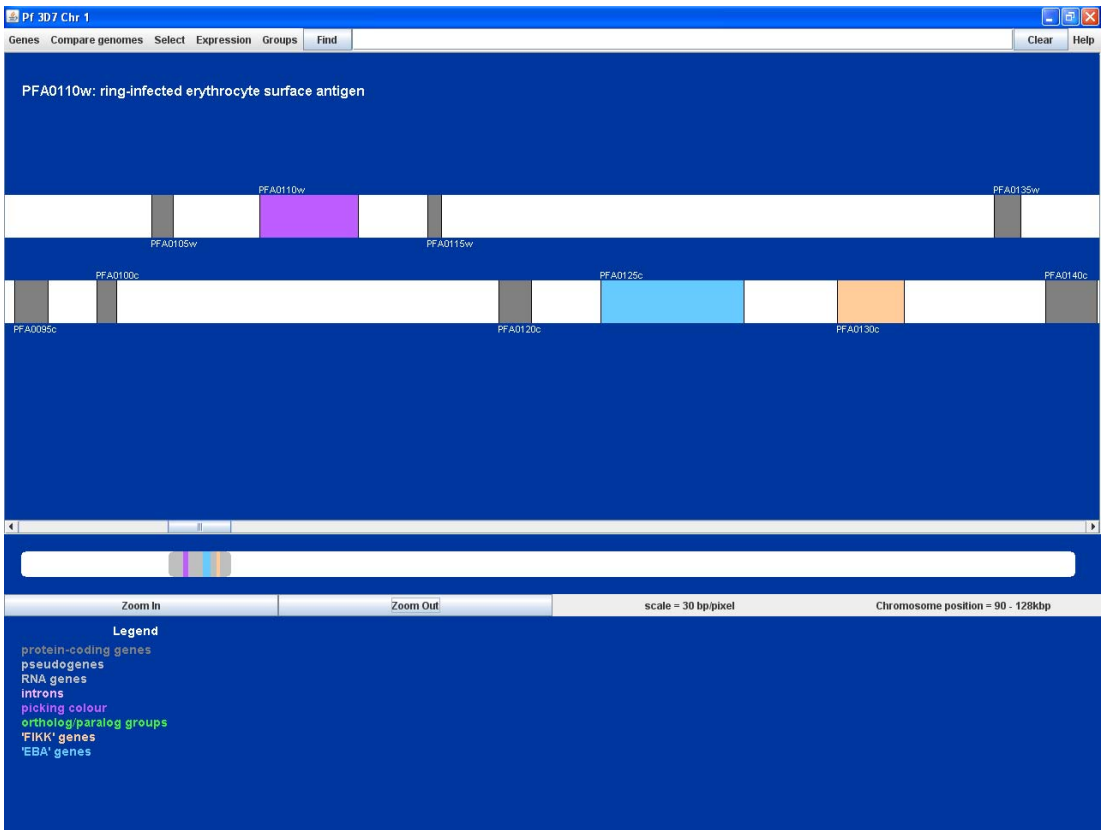


Figure 5.36. An example of tandem arrangement of subtelomeric multi-gene families. In this region on the left arm of chromosome 1 the genes include RESA (purple), EBA (blue) and FIKK kinase (orange).

5.3.2 Orthologs of FIKK kinases in other *Plasmodium* species

The FIKK kinase family has undergone the greatest expansion in *P. falciparum*. In most other sequenced *Plasmodium* genomes only a single copy has been found, including *P. berghei*, *P. vivax* and *P. knowlesi* (but not yet in *P. chabaudi*), as is demonstrated in the MaGnET ortholog data shown in Figure 5.37.

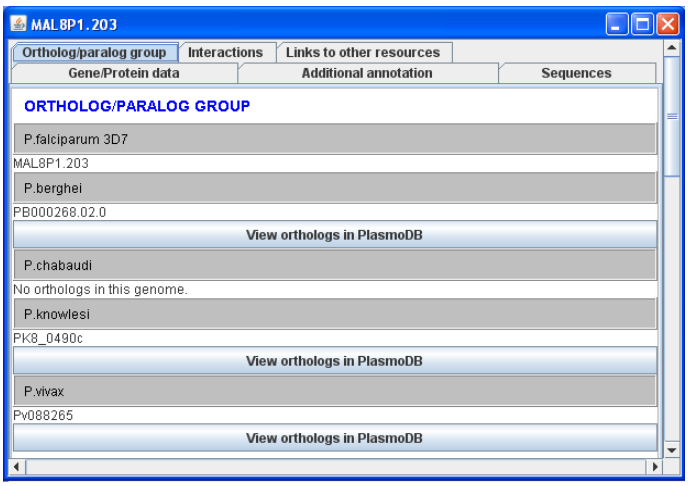


Figure 5.37. A single orthologous FIKK kinase gene is observed in most other *Plasmodium* species, including *P. berghei*, *P. knowlesi* and *P. vivax*, but not so far in *P. chabaudi*, probably due to low sequence coverage.

5.3.3 Differential expression of FIKK kinases

As Table 5.2 demonstrates, expression of the FIKK kinase paralogs is stage-specific and varies considerably across the asexual development stages. All the paralogs are expressed in at least one stage, at widely varying amplitudes and peaking at distinct time-points (Figures 5.38, 5.39 and 5.40). Figure 5.38 presents expression profiles for the genes with the greatest expression, including characterised trophozoite antigen R45 (PFD1175w), which is abundantly transcribed during ring and trophozoite stages. The genes in this group also include two of the three paralogs that were detected at gametocyte stages (Table 5.2). For the majority of paralogs, expression is quickly switched on and off, resulting in a short peak of expression (Figures 5.38, 5.39 and 5.40). Taken together, the results from microarray studies indicate that the FIKK kinase family does undergo differential expression in *P. falciparum*. This observation has been recently confirmed by a

study that showed that FIKK kinase expression changes in response to extra-cellular environmental factors (Nunes et al. 2007).

Gene	Sporozoite	Ring	Trophozoite	Schizont	Merozoite	Gametocyte
MAL7P1.144			X			X
PFA0130c			X		X	
PFE0045c				X	X	
PFL0040c		X	X	X	X	
MAL7P1.175			X			
PFI0095c	X		X	X	X	X
PFI0100c				X		
PFI0110c		X			X	
PFI0120c		X			X	
PFI0125c	X	X		X		
PFC0060c		X			X	
PF10_0160		X	X		X	
PF11_0510				X		
PF14_0733+4		X	X		X	
PF10_0380			X			
PFD1175w		X	X		X	X
PFD1165w		X	X		X	
PFI0105c	X	X	X			
PFI0115c			X		X	

Table 5.2. Life cycle stages where the *P. falciparum* 3D7 FIKK kinase genes were differentially expressed in microarray experiments (marked by an 'X'). Data are included for 19 of the 20 FIKK kinase paralogs because data for MAL8P1.203 were not available. The microarray expression data used to compile the table come from three datasets (Le Roch et al. 2003, Young et al. 2005 and Llinas et al. 2006) visualised through the MaGnET Expression Data Viewer. The gene expression information was visually compared between the datasets to decide based on consensus the life cycle stages where expression peaked.

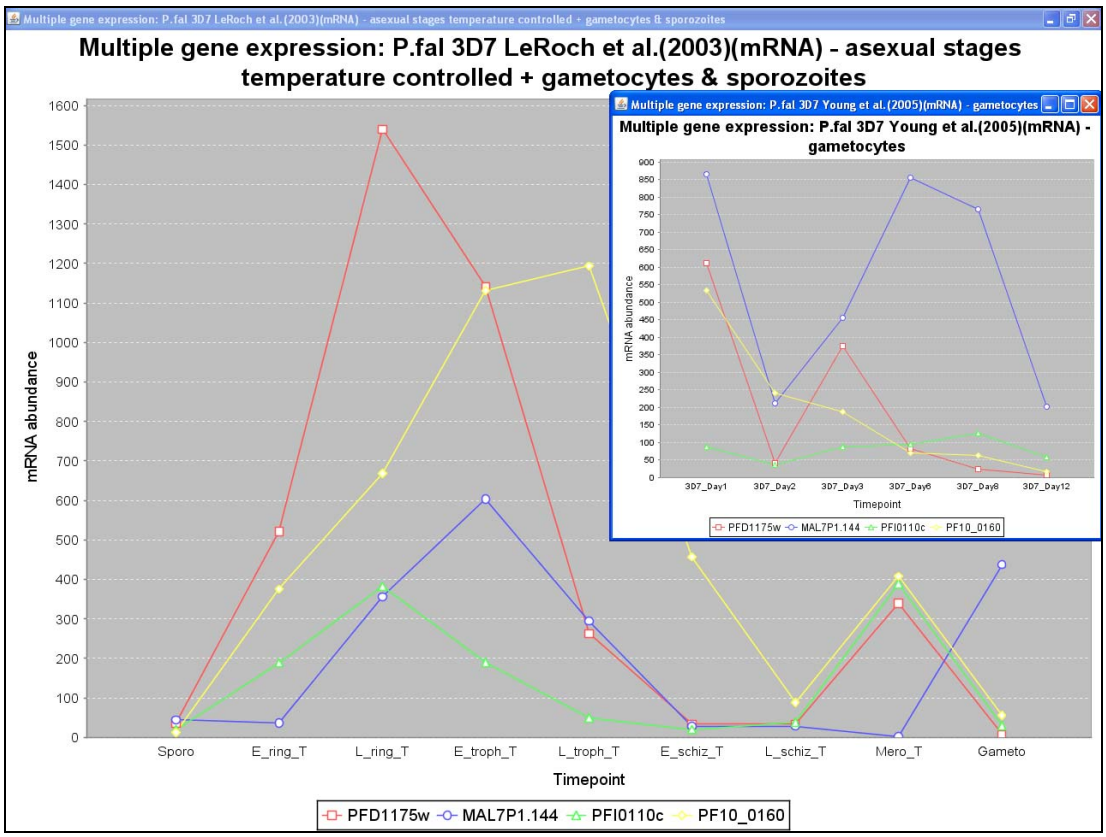


Figure 5.38. Time-series mRNA expression profile graphs for the highest expressed FIKK kinase paralogs, incorporating sporozoites, blood stages and gametocytes (inset) (3D7 strain data from Le Roch et al. 2003 and Young et al. 2005). The family member with highest mRNA abundance during the IDC is the R45 trophozoite antigen (PFD1175w), whereas in gametocytes the most abundantly transcribed member is MAL7P1.144.

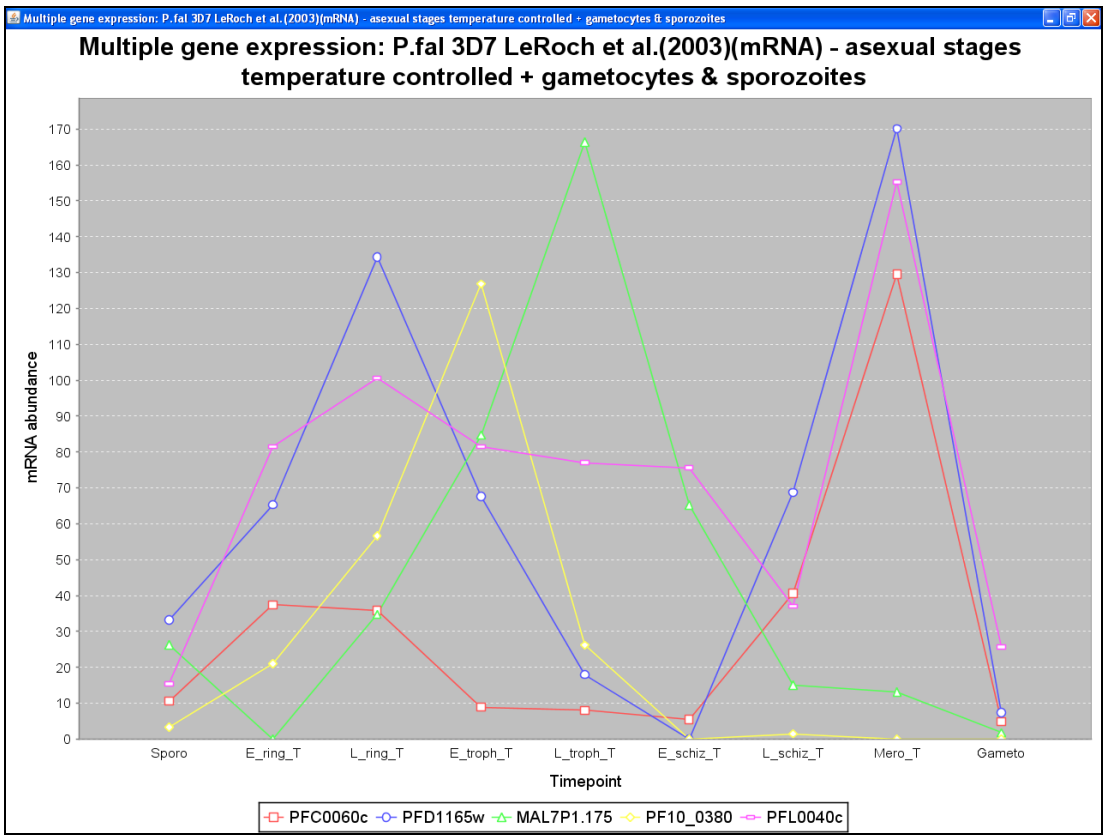


Figure 5.39. Time-series mRNA expression profile graphs for middle-range expressed FIKK kinase paralogs, incorporating sporozoites, blood stages and gametocytes (3D7 strain data from Le Roch et al. 2003).

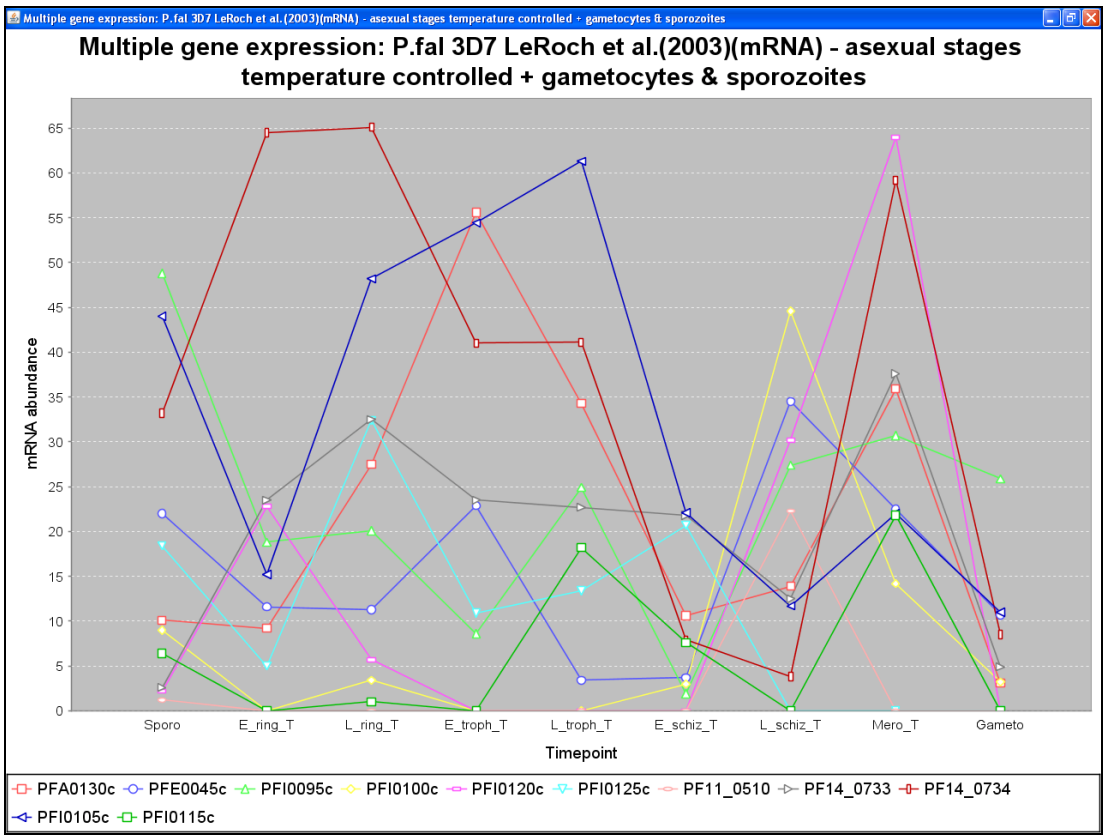


Figure 5.40. Time-series mRNA expression profile graphs for the lowest expressed FIKK kinase paralogs, incorporating sporozoites, blood stages and gametocytes (3D7 strain data from Le Roch et al. 2003).

Protein expression data indicated the presence of several of the FIKK kinase paralogs at the sporozoite stage and trophozoite stages (Figure 5.41). The R45 protein also appears to be one of the most abundant FIKK kinase proteins in blood stage parasites. Protein products of the internal stop codon-containing genes MAL7P1.175 and PF14_0733+4 were also detected at the sporozoite stage (Figure 5.41), which suggests that these proteins may be translated by read-through of the stop-codons. The cellular locations of some of these proteins as well as others have recently been revealed, with all but one being targeted to the IE (see Section 5.3.1.2) (Nunes et al. 2007).

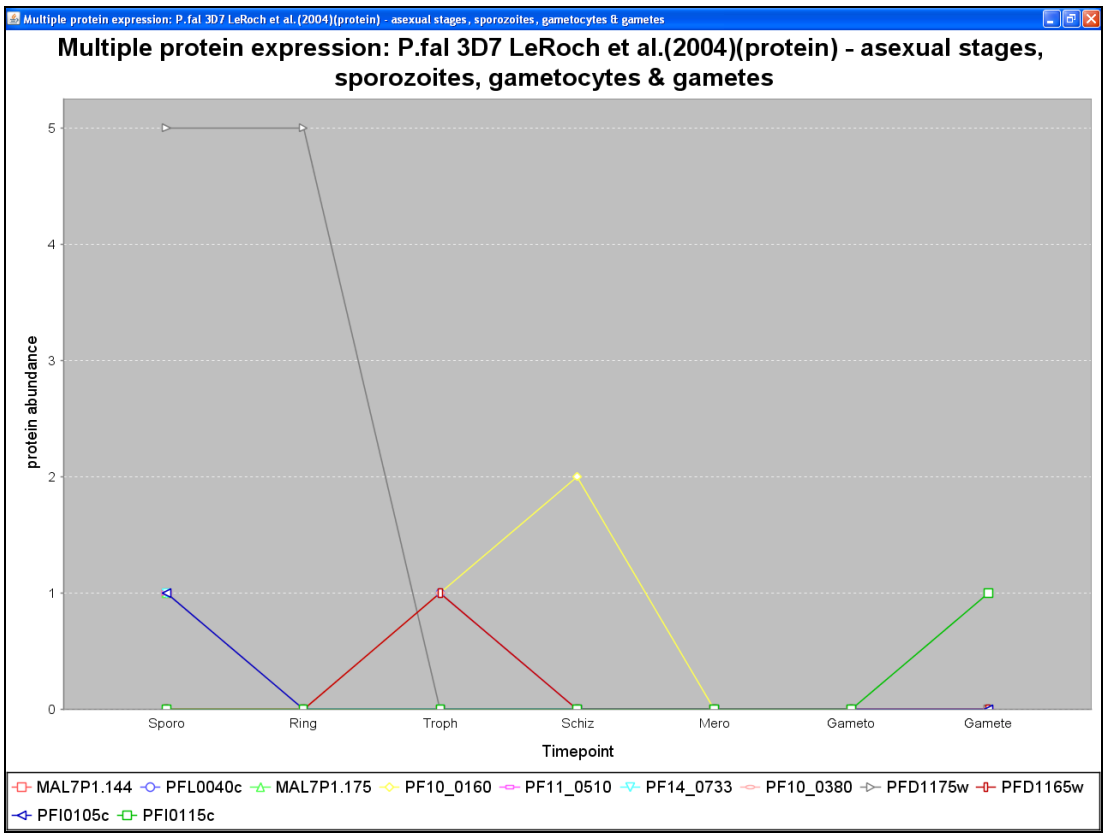


Figure 5.41. Time-series protein expression profiles for FIKK kinase paralogs. Various family members were recorded in sporozoites, blood stages, gametocytes and gametes (3D7 strain data from Florens et al. 2002 and Le Roch et al. 2004). The R45 trophozoite antigen (PFD1175w) was the most abundant family member in sporozoite and ring stages.

5.4 Discussion

This chapter presented three mini-studies describing how MaGnET can be used to demonstrate similar results to those from independent experimental and bioinformatic studies into gene function. They show that many of the observations that came from experimental characterisation of the time and location of protein expression are backed up by the expression data and other types of information (such as signal sequences) now made available through MaGnET. The majority of findings could be adequately demonstrated through MaGnET, with the main

exception being cases when the study used absolute expression levels from microarray data, where only relative expression data is available through MaGnET (see Section 5.1). Also, the cellular component annotation available within MaGnET is particularly incomplete due to lack of a systematic procedure to annotate the genes with GO terms when new reports appear characterising *Plasmodium* proteins, and no large-scale study of protein localisation has been completed. Therefore, little of the protein localisation results described in the original papers used in these mini-studies could be reproduced using MaGnET.

In some cases, the MaGnET analysis was able to go beyond the statements in the original papers to suggest new hypotheses on gene function (summarised in Table 5.3). The following sections discuss the strengths and weaknesses of MaGnET in context of the three studies.

Genes	Hypothesis
PFB0100c (KAHRP)	Expression of KAHRP mRNA can be detected in Dd2 and HB3 strain parasites, but may not be differentially expressed during the IDC to the extent of 3D7 parasites. Moreover, this gene is not absent or silenced at the transcription level in Dd2 parasites (as was thought), but it may not be translated or form a functional protein.
PFD1160w (SURFIN4.2)	This gene is located next to a region known to be involved in merozoite invasion pathway switching that is silenced in the Dd2 strain. As this gene is also silenced in Dd2, and the SURFIN4.2 protein is located in the merozoite apex region, it may also be involved in the process of invasion pathway switching.
PF13_0005 (RIFIN)	This *rifin* variant is differentially expressed in merozoites.
PFI1765c and PFI1770w	These genes are expressed and function in gametocytes.

Table 5.3. Summary of novel hypotheses about gene function that emerged from exploration of *P. falciparum* functional genomic data using MaGnET as described in Chapter 5.

5.4.1 The results of expression profiling of the IDC were successfully demonstrated using the MaGnET Expression Data Viewer

The microarray results of Llinas et al. brought to light several regions of the genome where differences in expression could be observed between *P. falciparum* strains. MaGnET was able to show through a series of individual gene expression profile graphs combining results from three strains that variations can occur for a number of reasons. In some cases, variations revealed differences in the timing of expression between strains; for example, an ATP binding cassette transporter protein that is expressed in different phases of the IDC in 3D7 and HB3 parasites. Many variations in recorded expression occurred due to high polymorphism of the region represented by the oligo, which is a common feature of surface antigens such as *Pf*EMP1, *Pf*EMP3, S-antigen, RESA-2 and KAHRP. The MaGnET Genome and Chromosome Viewers were also successfully used to demonstrate the exploration of differences in expression occurring over large regions, such as the putative deleted or silenced left arm of chromosome 2 in the Dd2 strain. MaGnET quickly facilitates the discovery of gene functions within the region of variable expression and the comparison of results from multiple expression studies.

In addition, the Query Builder facility within the Expression Data Viewer aided the discovery of a group of genes with large transcriptional variations between strains at a particular life cycle stage. This exercise proved useful in narrowing down a group of genes enriched for host-parasite interactions, such as cytoadherence and antigenic variation.

The Chromosome Viewer allows close-up examination of small regions of the genome where the genes are co-regulated and differ in expression between strains, which could have important functional implications. For example, MaGnET demonstrated variation in expression between 3D7 and HB3 parasites within a region of chromosome 4 containing genes linked to two different pathways for merozoite invasion, suggesting that different strains utilise different invasion pathways.

MaGnET was demonstrated to faciliate exploration of expression patterns of large multi-gene families of surface proteins, such as *var*, *rifin* and *stevor*. MaGnET effectively showed that parasite populations only utilise a small subset of the *var* genes and that even the laboratory strain 3D7 switches the subset of *var* genes it expresses. MaGnET exploration also revealed the presence of subsets of *rifins* that are differentially expressed at particular times in the life cycle. The different data viewers offered by MaGnET allow the evidence to be examined from various angles; for example, expression data was here viewed in the context of genomic location, across families and subsets and between datasets.

One area that MaGnET was not able to provide a demonstration of Llinas et al.'s results was in the indication of overall correlation between the expression patterns of the three strains. This could be made possible in future by the addition of pre-calculated correlation scores (such as Pearson correlation) for individual genes across strains. This data could also be used to enable searching for genes with correlated expression patterns, either within or across strains, based on a statistical measure of their similarity. This would be a useful feature for a future version of MaGnET to implement, because it is not currently offered by other tools in the field.

Llinas et al. also reported results of a comparison of the most highly expressed transcripts from their study with the most abundant proteins identified in two studies (Florens et al. 2002; Lasonder et al. 2002). The results of this comparison revealed that approximately 50% of the most abundant proteins were in the top 200 transcripts. Unfortunately, there was no way to reproduce this analysis in MaGnET because the data included only represents ratios of expression levels for individual oligonucleotides. Llinas et al. did not make the absolute expression levels available in a ready-to-use form that could easily be incorporated in the MaGnET database. This information could be extracted from the raw data files provided on the authors' website (Malaria IDC Comparison Database; http://malaria.ucsf.edu/comparison/index.php), but this would have required significant extra work and was not possible in the time-frame of this project.

The lack of absolute expression data for this dataset prevented some of the features the authors described in their paper from being demonstrated here. For instance, examination of expression data for multi-gene antigenic families was not completely repeatable using MaGnET because the actual expression level of the genes was not known. However, MaGnET was able to estimate which family members were being differentially expressed during the IDC by presenting a list of genes that underwent large changes in expression during the 48 hour cycle. Unfortunately, this approach did not take into account the presence of anomalies, such as extremely high or low recordings at single time-points.

Overall, MaGnET was able to provide access to all the resources used in the above study in one place. In addition to their own dataset the authors consulted three further mRNA and protein expression datasets (Florens et al. 2002; Lasonder et al.

2002; Le Roch et al. 2003), all of which are available through MaGnET. The authors also retrieved functional annotations, such as GO assignments, from PlasmoDB, which are also available from MaGnET, and, moreover, can easily be displayed in a helpful table format for a given list of genes.

5.4.2 MaGnET was used to explore a cluster of ring-stage exported proteins

MaGnET was able to effectively demonstrate results from a study by Spielmann et al. (2006) characterising the genes in a region of chromosome 9 where loss of functionality is linked to loss of cytoadherence. The MaGnET Chromosome Viewer was used successfully to reconstruct their analysis of existing transcription data to identify genes in this region that were specifically expressed at the ring stage. Four genes (whose products are termed the REX proteins) with putative functions in host-cell modification during the ring stage were selected for further characterisation. Visualisation of protein expression data with MaGnET showed that three out of the four REX proteins were detected in ring stages, which is consistent with Spielmann et al.'s results. Evidence of their being exported proteins was provided by signal/anchor sequence probabilities viewed from within MaGnET gene fact sheets.

MaGnET facilitated the exploration of intron/exon arrangement for the REX genes, revealing that they all have a similar two-exon arrangement. A combination of MaGnET functional genomic data, including domain predictions from InterPro and orthologous genes, agreed with the authors' conclusion that the REX proteins are not similar to any known proteins in other organisms, but that REX3 and REX4 are homologous to *P. vivax* and *P. knowlesi* genes.

Additionally, using MaGnET's methods for visualising expression data it was possible to distinguish between genes that had similar expression profiles to the REX genes but for which it was not safe to hypothesise that they could play a similar biological role. MaGnET provides the means for examining further evidence to back-up these theories; for example, Spielmann et al.'s suggestion that PFI1765c and PFI1770w induction in ring stages might arise from parasites committed to gametocytogenesis was corroborated using additional expression data for the gametocyte stages (data not shown) (Young et al. 2005).

In addition to their own protein expression data, Spielmann et al. consulted various genomic and functional data from PlasmoDB and transcription data from Le Roch et al. (2003). MaGnET offers the advantage that users can explore data from multiple mRNA and protein expression datasets in one place and quickly compare them. Exploratory analysis with MaGnET is supported by the provision of helpful functional annotation and, in some cases, predictions.

5.4.3 Many features of FIKK kinases were successfully demonstrated using MaGnET

The FIKK kinases are an interesting family of atypical protein kinases unique to *Apicomplexa*. The family's rapid expansion in *P. falciparum* and acquisition of unique properties, including signal sequences directing their export into the IE, have raised interesting questions about their possible role as signalling mediators between host and parasite (Schneider and Mercereau-Puijalon 2005; Nunes et al. 2007). In this mini-study many aspects of *P. falciparum* FIKK kinase paralog genomic organisation, protein structure and expression were demonstrated using various MaGnET features.

The MaGnET Genome Viewer could be used to rapidly display the genomic location of the paralogs, revealing their typical subtelomeric localisation. The Chromosome Viewer intron display capability allowed the intron/exon arrangement to be examined, which revealed a typical arrangement of a small exon 1, long exon 2 and short exon 3. It also encouraged the discovery of a few anomalies, such as a case of probable gene misprediction where two annotated genes represent just one paralog.

Various MaGnET tools including the informative gene fact sheets and the Data Analysis search facility were used to discover the structural and functional features the proteins had in common. Helpful available annotation included comparatively modelled structures for the C-terminal region of many of the proteins, indicating their similarity to kinase catalytic domains. Furthermore, the inclusion of InterPro predicted domains and sequence features confirmed the universal presence of kinase-like domains and motifs, as well as the common occurrence of N-terminal hydrophobic sequences predicted as transmembrane or signal/anchor sequences.

Evidence of higher-order genomic arrangements of subtelomeric multi-gene families in *P. falciparum* was presented by the often-time association of FIKK kinases with members of families such as DNA J domain protein-encoding genes (including RESA), EBA, fatty acid CoA synthase and various hypothetical protein-encoding gene families.

The MaGnET Expression Data Viewer offered the opportunity to explore the transcription patterns of the family as recorded in various microarray experiments. Individual FIKK kinase paralogs are clearly differentially expressed and function at specific stages of the life cycle. Moreover, the expression profiles of many genes

showed sharp peaks at individual stages, indicating a directed mechanism for switching transcription on and off. Protein expression data for several family members at multiple life cycle stages is further evidence of their varied functions during asexual development. Interestingly, mRNA and protein expression data indicated that the most abundant member of the family is the trophozoite antigen R45, a known surface antigen that is unique amongst the family due to a large inserted repeat sequence.

The MaGnET ortholog/paralog data showed that there appear to be only single copies in other *Plasmodium* species. If ortholog information becomes available for *P. reichenowi*, which is the most similar to *P. falciparum* and also has several paralogs, it would be useful to provide more detailed comparisons between ortholog pairs.

An improvement to the existing MaGnET functionality that would help in a similar study would be the addition of structure classification data, as discussed in Chapter 4 (if indeed it is possible to attain this information automatically). Structural domain classification for the comparative models would provide useful hints to malaria researchers who are not familiar with common structural motifs, such as the kinase domain.

To complete their study Schneider and Mercereau-Puijalon consulted two transcription datasets (Bozdech et al. 2003; Le Roch et al. 2003), protein expression data (Florens et al. 2002), genome maps from PlasmoDB and the *P. falciparum* 3D7 genome publication, PlasmoDB and GeneDB for orthologs and annotation. All this information is available in one place in MaGnET via the integrated viewers and local database combining data from multiple sources.

5.4.4 Limitations of MaGnET for functional genomic data analysis

Although the majority of results from the publications described in this chapter could be demonstrated successfully with MaGnET, there remain some areas that could not be covered. Aside from results of specialised laboratory experiments, such as protein localisation studies (discussed above), that MaGnET could not reproduce, there are limits to the extent of functional genomic data analysis that MaGnET is capable of. Specific functionality that MaGnET lacks is interactive sequence analysis, which was used by the publications described in this chapter for identifying gene families and examining their homology to other sequences.

Future development of MaGnET could provide integrated tools for sequence analysis. For example, it would be useful if MaGnET linked to a sequence homology search tool, such as BLAST (Altschul et al. 1990). This would be important for searching proteins in other organisms for similarities, but also for internal searching within the *P. falciparum* genome. Such functionality becomes even more important once other *Plasmodium* species are added. It would also be helpful to facilitate users to perform sequence alignments of selected genes or proteins. This would have a number of applications, such as exploring conserved functional sites in protein families. While PlasmoDB provides tools for BLAST searching within *Plasmodium* species, it does not provide opportunities for searching other organism data and does not provide sequence alignment tools.

6. HYPOTHESIS GENERATION THROUGH EXPLORATION USING MAGNET

Overview

This chapter will present interesting avenues of research that have developed from exploration of functional genomic data using MaGnET. Along the way clues were picked up about probable gene function, including many uncharacterised genes, which were pieced together into plausible, testable hypotheses. Further lines of evidence are provided through the careful use of complementary resources, including statistical data and relevant literature. The likely significance of the findings in the context of the field will be discussed. At the end of the chapter all the predictions will be summarised and follow-up experiments to test them, suggested.

6.1 P. falciparum cyclin-dependent kinases and their cyclin partners

Cyclin-dependent kinases (CDKs) are major regulators of eukaryotic cell cycle progression. CDKs remain inactive until they bind to their cognate cyclin molecules. The active complexes phosphorylate a number of proteins involved in processes required for cell division, such as DNA synthesis and chromosome segregation. In the metazoan and yeast model of cell cycle control, individual cyclins have a narrow window of expression compared to CDK subunits, whose expression is not as rigorously regulated. Each cyclin only binds a subset of CDKs, and vice versa. In this way, the different enzymes will become active only at certain

times during the cycle, regulating transition between different phases [for a review of eukaryotic cell cycle regulation see (McGowan 2003)]. Several CDKs are found in mammalian cells; however, only a few directly regulate cell cycle processes. Others are involved in regulation of transcription or neuronal functions. Similarly, many cyclins also exist.

Malaria parasites undergo several rounds of replication during their life cycle, including mitosis and meiosis within the mosquito stages, pre-erythrocytic schizogony within hepatocytes and erythrocytic schizogony. Little is known about cell cycle processes and control mechanisms in *Plasmodium*. Nevertheless, it is clear that the cell cycle differs significantly from the metazoan/yeast model (Arnot and Gull 1998; Doerig 2005). In the classical cell cycle model, four distinct phases occur: G_1, where the cell grows and builds resources, S phase, where DNA is replicated once, G_2, preparation for cell division, and M, where the genomes are segregated and the cell divides. In erythrocytic malaria parasites (the replicative stage most studied), the phases are not so neatly defined. The merozoite and ring stage most likely correspond to G_1 phase, and S phase appears to begin about 18 hours after invasion. However, after DNA replication begins the cell cycle can no longer be described according to the traditional model due to several asynchronous nuclear divisions occurring within a single schizont (Arnot and Gull 1998; Doerig 2005). This makes cell cycle progression very difficult to study and leaves open many questions as to how the cell achieves asynchronous nuclear division when all nuclei are subject to the same cytoplasmic conditions.

The work described in the following section will review the current status of known and predicted CDK and cyclin homologues in *P. falciparum* 3D7 and

demonstrates the use of MaGnET to investigate properties of the genes. The analysis reveals distinct patterns of CDK-cyclin expression during erythrocytic stages and leads to hypotheses regarding possible subunit pairings at various cell cycle phases.

6.1.1 Cyclin-dependent kinases and related proteins in *P. falciparum*

To date, seven genes in the *P. falciparum* genome that code for cyclin-dependent kinases (CDKs) and CDK-related kinases (CRKs) have been characterised (Table 6.1) [reviewed in (Doerig et al. 2002; Ward et al. 2004)]. Ward et al. (2004) used phylogenetic analysis to further predict CDK function for one uncharacterised gene, MAL13P1.196.

Gene	CDK/CRK	Cyclin-dependency	Reference
MAL13P1.279	*Pf*PK5	Cyclin-dependent	(Ross-Macdonald et al. 1994)
MAL13P1.185	*Pf*PK6	Cyclin-independent	(Bracchi-Ricard et al. 2000)
PF10_0141	*Pf*mrk	Cyclin-dependent	(Li et al. 1996)
PFD0865c	*Pf*crk-1	No data	(Doerig et al. 1995)
PFD0740w	*Pf*crk-3	No data	(Doerig et al. 2002)
PFC0755c	*Pf*crk-4	No data	(Doerig et al. 2002)
PFF0750w	*Pf*crk-5	No data	(Ward et al. 2004)
MAL13P1.196	–	No data	(Ward et al. 2004)

Table 6.1. CDKs and CRKs of *P. falciparum*. Adapted from (Doerig et al. 2002).

6.1.2 *P. falciparum* cyclins

The first *P. falciparum* cyclin (*Pf*cyc-1; PF14_0605), an apparent ortholog of mammalian cyclin H, was discovered by Le Roch et al. (2000) after searching an early release of initial genome sequence constructs. Following completion of the genome sequence, Merckx et al. (2003) were able to identify and characterise three

further cyclins: *Pf*cyc-2, *Pf*cyc-3 and *Pf*cyc-4 (PFL1330c, PFE0920c and PF13_0022, respectively).

6.1.3 CDK-cyclin combinations

In mammals, the usual partner for cyclin H is CDK7, for which it is highly specific. *Pf*cyc-1 (an ortholog of cyclin H) was shown to activate *Pf*PK5 *in vitro*, which was surprising since *Pf*PK5 is an ortholog of CDK1 and not CDK7 (Le Roch et al. 2000). Additionally, *Pf*PK5 was shown to be activated by human p25, a specific activator of human CDK5 that despite displaying no sequence homology to cyclins has a similar tertiary structure (Le Roch et al. 2000).

Of the other *P. falciparum* cyclins, *Pf*cyc-3 was shown to potently activate *Pf*PK5, with *Pf*cyc-4 marginally activating it and *Pf*cyc-2 showing no activity *in vitro*. Interestingly, in the presence of *Pf*cyc-3, *Pf*PK5 is not able to autophosphorylate, but it can with *Pf*cyc-1 (Merckx et al. 2003). Further evidence of *Pf*PK5's promiscuity is provided by the ability of RINGO (a *Xenopus* protein that can activate CDKs despite no homology to known cyclins) to activate it more strongly than any *P. falciparum* cyclin (Merckx et al. 2003). Since *Pf*PK5 displays an unusual ability to be activated by multiple cyclins, including those that do not have sequence similarity to traditional cyclins, it leaves open the possibility of novel cyclin-like proteins within the *P. falciparum* genome that are undetectable by sequence searches.

The plasmodial orthologue of CDK7 is *Pf*mrk. CDK7 in mammals has dual functions as a transcription regulator through phosphorylation of RNA polymerase II and a CDK-activating kinase (CAK). CAK activity is dependent on binding of

cyclin H. When MAT1 joins the complex, it is able to regulate transcription through RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. So far no evidence has been found for CAK activity of *Pf*mrk. Both *Pf*cyc-1 and cyclin H activate *Pf*mrk-mediated CTD phosphorylation. *Pf*MAT1 (PFE0610c) stimulates this activity in a cyclin-dependent manner (Chen et al. 2006b).

6.1.4 Retrieval of further CDKs, cyclins and associated proteins

The predicted annotation data available within MaGnET (GO and InterPro predictions based on sequence similarity) was queried for additional hits to CDK and cyclin sequences not characterised thus far. Searches of the MaGnET database using the Data Analysis search facility revealed no further predicted CDKs or CRKs. A search for proteins with similarity to cyclins led to a list of four novel proteins possessing cyclin-like domains (Table 6.2).

Gene	Product name	Domains	E-value of match
PFF0270c	Cyclin dependent kinase binding protein	Cyclin-like; Cdk5 and c-Abl linker protein cables	$4.3E^{-15}$; 0
MAL8P1.152	Hypothetical protein	Cyclin-like	$2.9E^{-10}$
PF10_0139	Hypothetical protein	Cyclin-like	$2E^{-9}$
MAL13P1.131	Hypothetical protein	Cyclin-like	$2.1E^{-7}$

Table 6.2. List of proteins with predicted cyclin-like domains from InterPro annotation.

It should be noted here that possession of a cyclin-like domain is not sufficient for indisputable cyclin function because cyclin-like domains are also found in other proteins. For example, two other *P. falciparum* proteins that also have InterPro-annotated cyclin-like domains are putative transcription factors (PF14_0469 and PFA0525w, data not shown). Some transcription factors are known to possess

cyclin-like domains, but they do not regulate CDKs (Noble et al. 1997). This leaves open the question about whether the proteins in Table 6.2 with cyclin-like domains function either as CDK-regulating cyclins or as transcription factors.

Investigations using the MaGnET Data Analysis Viewer for other CDK-associated proteins turned up characterised *Pf*MAT1 and also a putative S-phase kinase-associated protein 1 (Skp1) (MAL13P1.337), which possibly functions as a CDK/cyclin-associated protein.

6.1.5 Using expression data to predict likely *in vivo* CDK/cyclin complexes

The case discussed above of *Pf*PK5's unusual *in vitro* activation by several different cyclins illustrates how difficult it is to predict functional CDK/cyclin complexes from sequence information alone. Furthermore, even if a particular cyclin is able to activate a CDK *in vitro*, the pairing may not occur *in vivo* due to other factors, such as timing of expression of the components, presence of inhibitors, binding of co-activators that enhance the stability of alternative CDK/cyclin complexes, so it may have no functional relevance. Data about one of the aspects determining functional CDK/cyclin pairs, namely co-expression of the components, could provide a useful starting point for narrowing down the possible *in vivo* combinations. The aim of this "mini-study" is to use available expression data to predict co-expressed pairs of *P. falciparum* CDKs and cyclins that may form functional complexes.

The recorded erythrocytic stage expression profiles of all pairwise combinations of known and predicted *P. falciparum* CDKs and cyclins were

compared in three time-course experiments involving 3D7, Dd2 and HB3 strain parasites (Bozdech et al. 2003; Llinas et al. 2006). Close similarity in timing and amplitude of expression profiles were examined and assessed visually using the MaGnET Expression Data Viewer (Table 6.3) and assertions made about the probable correlation between pairs or groups of CDKs and cyclins.

The results show that some of the CDK-encoding genes appear to have highly similar expression profiles to several of the putative cyclins and vice versa. Moreover, a few CDK/cyclin gene pairs' expression profiles seem to be highly correlated to each other and not to other genes in Table 6.3. Details are provided for some examples in the following section.

		MAL13P1.185	MAL13P1.196	PFC0755c[$]	PFD0865c	PFD0740w	PF10_0141	MAL13P1.279	PFF0750w
		***Pf*PK6**	**-**	***Pf*crk-4**	***Pf*crk-1**	***Pf*crk-3**	***Pf*mrk**	***Pf*PK5**	***Pf*crk-5**
PF13_0022	***Pf*cyc-4**	Similar	X	X	Similar; especially in HB3	X	X	Yes	X
PF14_0605	***Pf*cyc-1**	Similar; least in 3D7	X	X	X	Similar in 3D7	Yes	X	X
PFE0920c[#]	***Pf*cyc-3**	X	X	X	X	X	X	X	X
PFL1330c	***Pf*cyc-2**	X	Yes	Yes	Similar; PFL1330c few hrs behind	X	X	Similar; PFL1330c few hrs behind	Yes
PFF0270c*	**CDK binding**	X	Yes	Yes; after hr 24 *Pf*crk4 ~2 hrs behind	Similar; PFF0270c ~3 hrs behind	X	X	Similar	Yes; especially in HB3
MAL8P1.152	**Hypothetical**	2nd half of IDC is similar	X	X	X	HB3 very similar; 3D7 similar	Yes; especially in Dd2	X	X
PF10_0139	**Hypothetical**	Similar	X	X	X	Similar; especially in 3D7	Similar; especially in 3D7	X	X
MAL13P1.131	**Hypothetical**	X	Yes; least in Dd2	Yes; PFC0755c slightly behind	Similar; especially in Dd2	X	X	Similar	Similar
MAL13P1.337	**Skp1 family**	Similar; MAL13P1.337 few hrs behind	X	X	Yes	X	X	Similar	Similar; PFF0750w few hrs behind

[$] based on HB3 only

[#] based on HB3 & 3D7 only

* based on HB3 & Dd2 only

Yes = same overall shape and close match between individual data points

Similar = same overall shape with some variation between individual data points

X = no similarity

Table 6.3. Comparison of CDK and cyclin expression profiles in the IDC [data for *P. falciparum* strains 3D7 and Dd2 from Llinas et al. 2006 and HB3 from Bozdech et al. 2003]. Note: the comparisons were made entirely visually, so are subjective rather than objective scores of similarity. It is included as an example of a process a user might go through will making initial investigations using MaGnET. Any observations made about similarity of expression profiles should be confirmed using statistical measures before laboratory experiments are undertaken.

6.1.5.1 The components of the RNA polymerase II CTD phosphorylation complex, Pfmrk, Pfcyc-1 and PfMAT1, have highly similar expression profiles

To test whether the components of the characterised CTD phosphorylation complex, *Pf*mrk, *Pf*cyc-1 and *Pf*MAT1 (Chen et al. 2006b), have similar expression profiles, time-series graphs were created of their expression profiles during the IDC. Figure 6.1 shows their expression profiles in the Dd2 strain. Their expression peaks in the early-mid trophozoite stage (hours 14-25), during the period of rapidly increasing RNA synthesis and parasite growth (Gritzmacher and Reese 1984). The synchronous expression of all three components during the time-frame for initiation of RNA synthesis reflects the role of the *Pf*mrk/*Pf*cyc-1/*Pf*Mat1 complex as a transcription regulator.

Comparison of the expression profiles of the *Pf*mrk, *Pf*cyc-1, *Pf*MAT1 and *Pf*PK5 demonstrates that *Pf*PK5 has a distinctly different expression profile to the other three genes (Figure 6.2). The expression profile of *Pf*PK5 dips during the ring/early trophozoite stages (hours 3-18) when the *Pf*mrk/*Pf*cyc-1/*Pf*Mat1 complex expression rises, and peaks in the late trophozoite/early schizont stages (hours 21-35). The marked difference in their expression profiles supports the *in vitro* observation that *Pf*PK5 is not a substrate for CAK activity of the *Pf*mrk/*Pf*cyc-1/*Pf*Mat1 complex (Chen et al. 2006b).

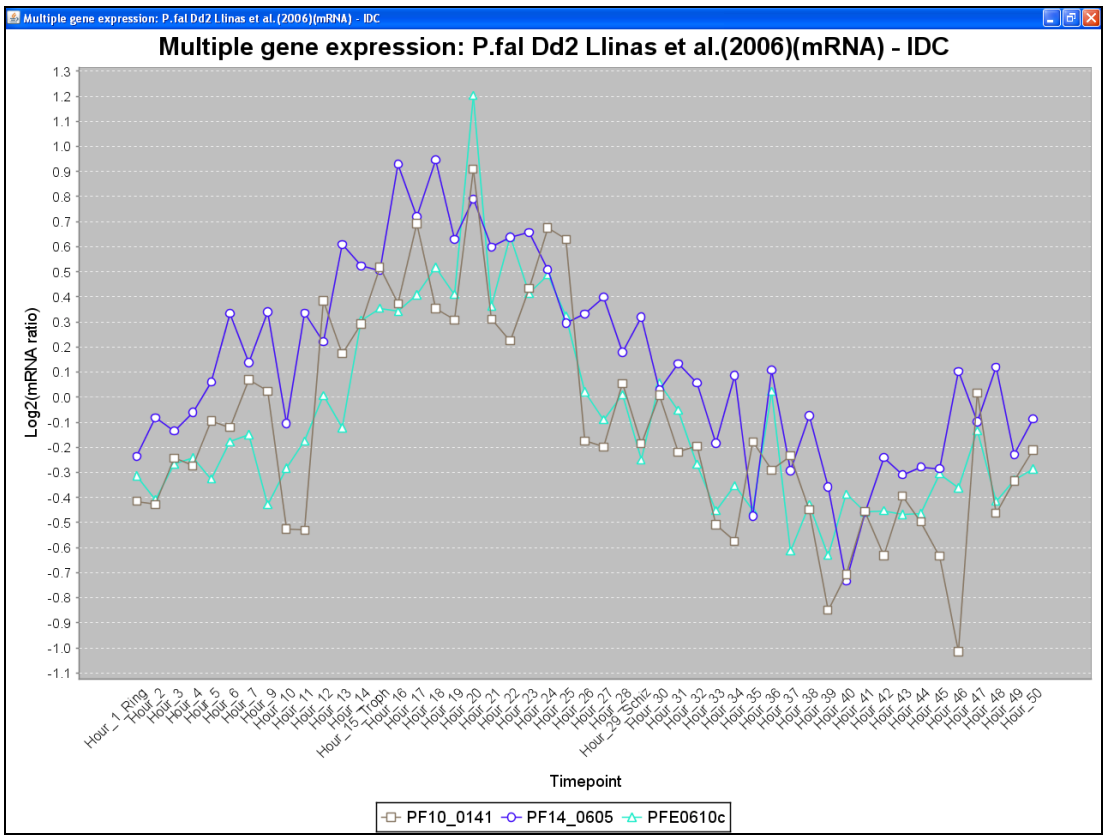


Figure 6.1. Time-series graph of expression of the genes *Pf*mrk (PF10_0141), *Pf*cyc-1 (PF14_0605) and *Pf*MAT1 (PFE0610c) during the Dd2 IDC (data from Llinas et al. 2006).

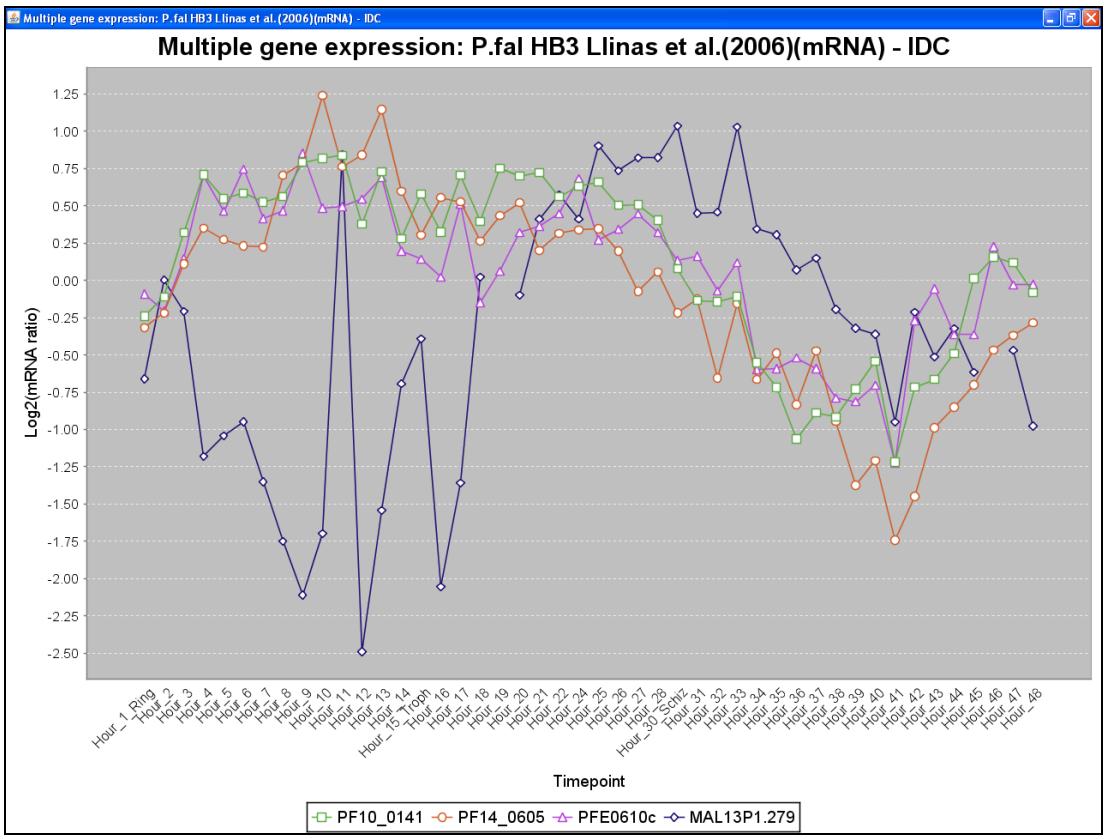


Figure 6.2. Time-series graph of expression of the *Pf*mrk/*Pf*cyc-1/*Pf*Mat1 complex (encoded by genes PF10_0141, PF14_0605 and PFE0610c) and *Pf*PK5 (MAL13P1.279) during the HB3 IDC (data from Bozdech et al. 2003).

6.1.5.2 PfPK5 has a similar expression profile to Pfcyc-4 and Pfcyc-2 but not Pfcyc-1 and Pfcyc-3

The only other characterisation study of a *P. falciparum* CDK to date showed that *Pf*PK5 can be activated strongly *in vitro* by *Pf*cyc-3 and less so by *Pf*cyc-1 and *Pf*cyc-4 (Merckx et al. 2003). Comparison of the IDC expression profiles of *Pf*PK5 with all the known and predicted cyclins (Table 6.3) showed it had a highly similar profile to *Pf*cyc-4 (Figure 6.3). Additionally, *Pf*cyc-2 has a reasonably similar profile to *Pf*PK5, but it peaks later, during the schizont stage (hours 28-49), whereas *Pf*PK5 and *Pf*cyc-4 peak earlier, during the late trophozoite/early schizont stages

(hours 21-36) (Figure 6.3). There is no similarity between the profiles of *Pf*PK5 and either *Pf*cyc-1 or *Pf*cyc-3 during this part of the life cycle (Figure 6.4). Since the parasite undergoes several different rounds of cell division during its life cycle (Arnot and Gull 1998), it is feasible that *Pf*PK5 forms functionally distinct complexes with different cyclins at different stages.

To investigate this further the expression profile of *Pf*PK5 was compared to those of cyclins 1-4 during the gametocyte stage of the life cycle (Figure 6.5). *Pf*PK5 and *Pf*cyc-2 are both expressed during day two of development, after which their expression tails off.

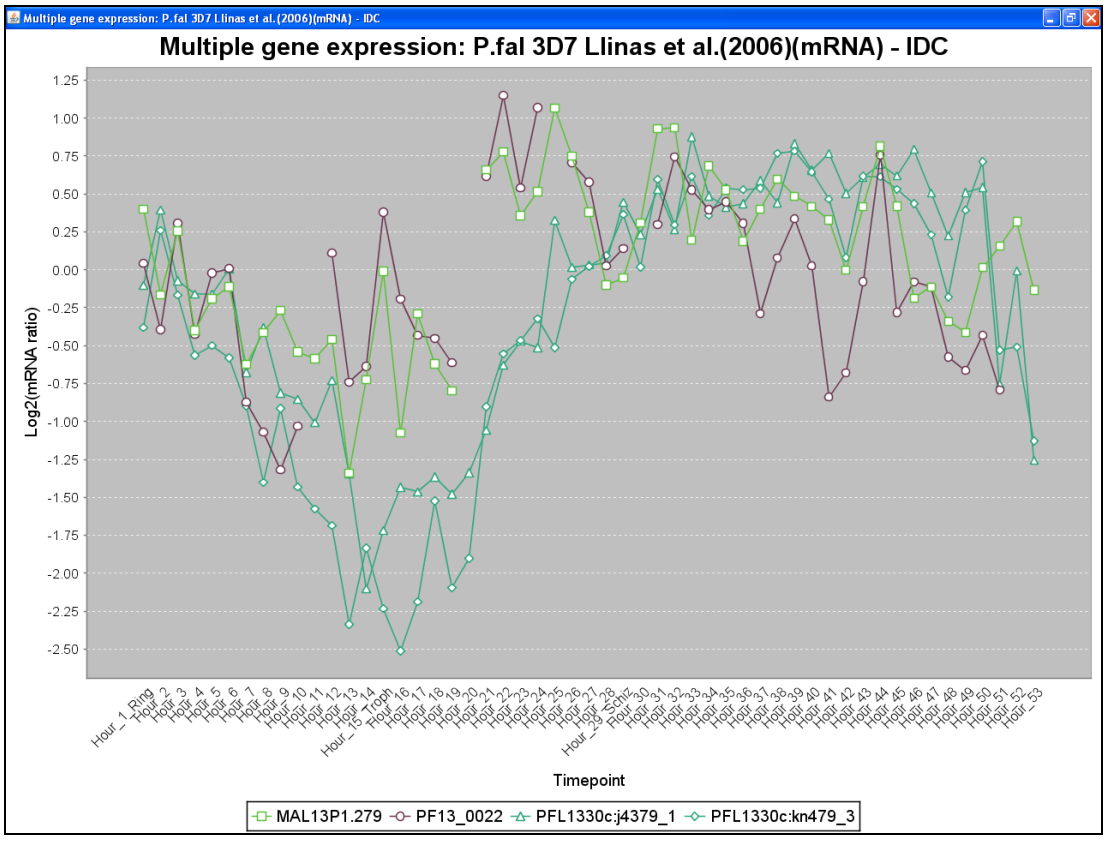


Figure 6.3. Time-series expression profiles of *Pf*PK5 (MAL13P1.279), *Pf*cyc-2 (PFL1330c) and *Pf*cyc-4 (PF13_0022) during the 3D7 IDC (data from Llinas et al. 2006).

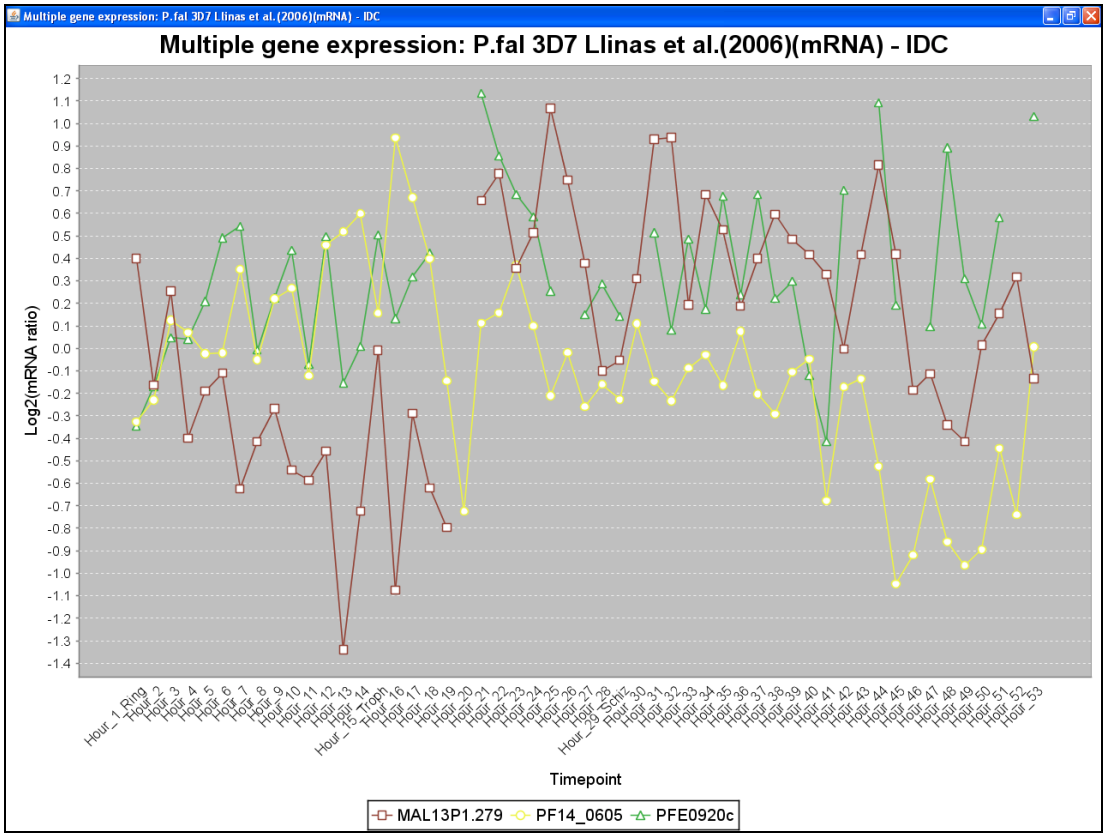


Figure 6.4. Time-series expression profiles for *Pf*PK5 (MAL13P1.279), *Pf*cyc-1 (PF14_0605) and *Pf*cyc-3 (PFE0920c) during the 3D7 IDC (data from Llinas et al. 2006).

From these results it is tempting to speculate that *Pf*PK5 may form a functional complex with *Pf*cyc-4 during the trophozoite stage of the IDC. The lack of *in vitro* activity recorded for *Pf*PK5/*Pf*cyc-2 (Merckx et al. 2003) does not rule out the possibility of *in vivo* activity. The functional complex may require the presence of a co-activator protein, perhaps to enhance stability, and this protein could, for example, be present during the gametocyte stage.

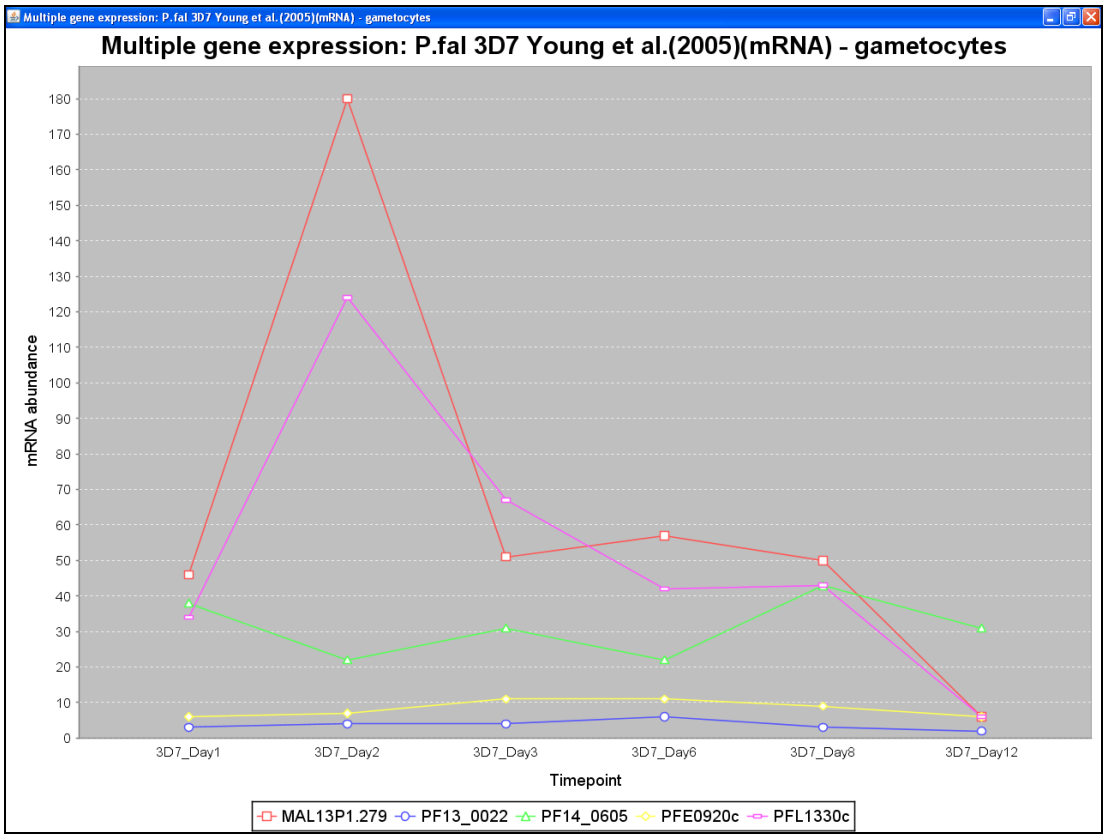


Figure 6.5. Time-series expression profiles of *Pf*PK5 (MAL13P1.279) and *Pf*cyc1-4 (PF14_0605, PFL1330c, PFE0920c and PF13_0022) in 3D7 gametocytes (data from Young et al. 2005). Both *Pf*PK5 and *Pf*cyc-2 are expressed during day two of gametocyte development.

6.1.5.3 A group of three CDKs and three cyclins are co-expressed in schizonts

A pattern emerging from Table 6.3 is the consistent similarity between the expression profiles of one group of three CDKs and three cyclins. The CDKs are *Pf*crk-4, *Pf*crk-5 and a putative, unnamed CDK/CRK (MAL13P1.196) that clusters phylogenetically with *Pf*crk-4 (Ward et al. 2004). The cyclins include *Pf*cyc-2 and two putative cyclin-like proteins, PFF0280c and MAL13P1.131. Figure 6.6 shows the highly similar expression profiles of all six genes during the IDC. Their

expression increases during the late trophozoite stage, peaking in the early schizont (hours 22-40).

From this data alone it is difficult to pick out possible pairs of interacting CDK/cyclins. However, one can speculate that all six CDKs and cyclins will be involved in regulating the same process within schizonts. According to a model of the *P. falciparum* erythrocytic cell cycle put forward by Leete and Rubin in 1996, the schizont stage starts with a series of rapid rounds of DNA synthesis and nuclear mitosis. The number of nuclei produced in each schizont is variable from 8 to 26, which suggests that the nuclei are not progressing synchronously through the cell-cycle. To reconcile this with the usual synchronised behaviour of nuclei under these conditions where regulation of cell cycle is driven by waves of cyclin expression, a model was proposed whereby CDK/cyclin complexes involved in both DNA replication and mitosis must exist in abundance in the schizont cytoplasm and there exists a mechanism for maintaining cell cycle integrity within each nucleus (Leete and Rubin 1996). The fact that schizont nuclei keep their membranes intact during nuclear division provides a means to regulate their cyclin content separately from the pool of cyclins in the cytoplasm. By selective import and degradation of cyclins from the cytoplasm the nuclei can individually regulate their cell cycle. Leete and Rubin proposed that as long as cyclins remain above a threshold level in the cytoplasm each nucleus will continue to initiate new rounds of mitosis.

The closely regulated expression of a set of three CDKs and three cyclins in the early schizont certainly fits with the model of a pool of CDK/cyclins within the cytoplasm that can be selectively imported by each nucleus, as required. The rapid

switch on and off of these six genes demonstrated in Figure 6.6 would provide a method for regulation of availability of CDK/cyclins in the cytoplasmic pool.

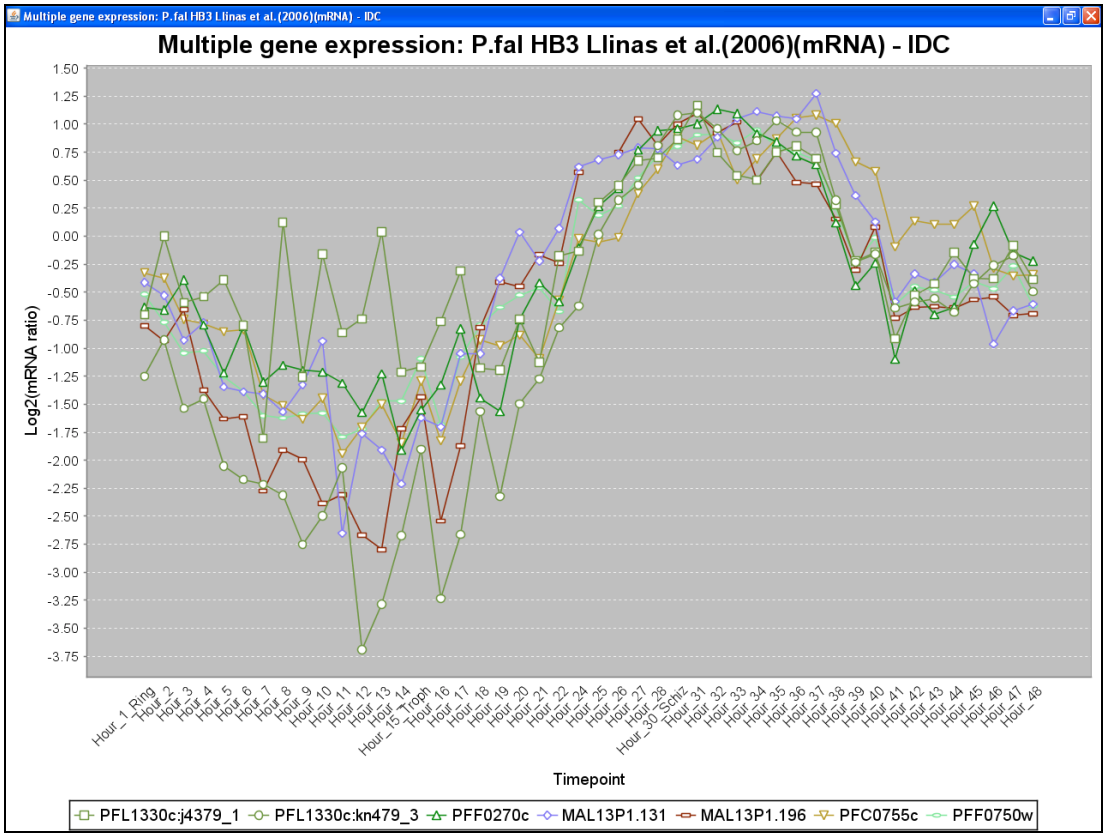


Figure 6.6. Time-series expression profiles of three putative cyclins [*Pf*cyc-2 (PFL1330c), PFF0270c and MAL13P1.131] and three CDKs [*Pf*crk4 (PFC0755c), *Pf*crk5 (PFF0750w) and MAL13P1.196] during the HB3 IDC (data from Bozdech et al. 2003).

6.1.5.4 A second group of three CDKs and three cyclins are co-expressed during the ring and trophozoite stages

A second group of three CDKs and three putative cyclins were observed to follow similar expression profiles, with peak expression occurring during the late ring/early trophozoite stages (hours 4-26) (Figure 6.7). Here, the CDKs include *Pf*crk3, *Pf*mrk and *Pf*PK6 and the potential cyclins are the hypothetical proteins

PF10_0139, MAL8P1.152 and the *Pf*mrk-activator *Pf*cyc-1. Although there is a large body of evidence to suggest that *Pf*PK6 does not require a cyclin to be active, there remains open the possibility that binding to a cyclin will further increase its activity (Bracchi-Ricard et al. 2000). As above, it is difficult to determine from this data alone which cyclin might pair with *Pf*crk-3, although MAL8P1.152 is noted to have a particularly similar profile to *Pf*crk-3.

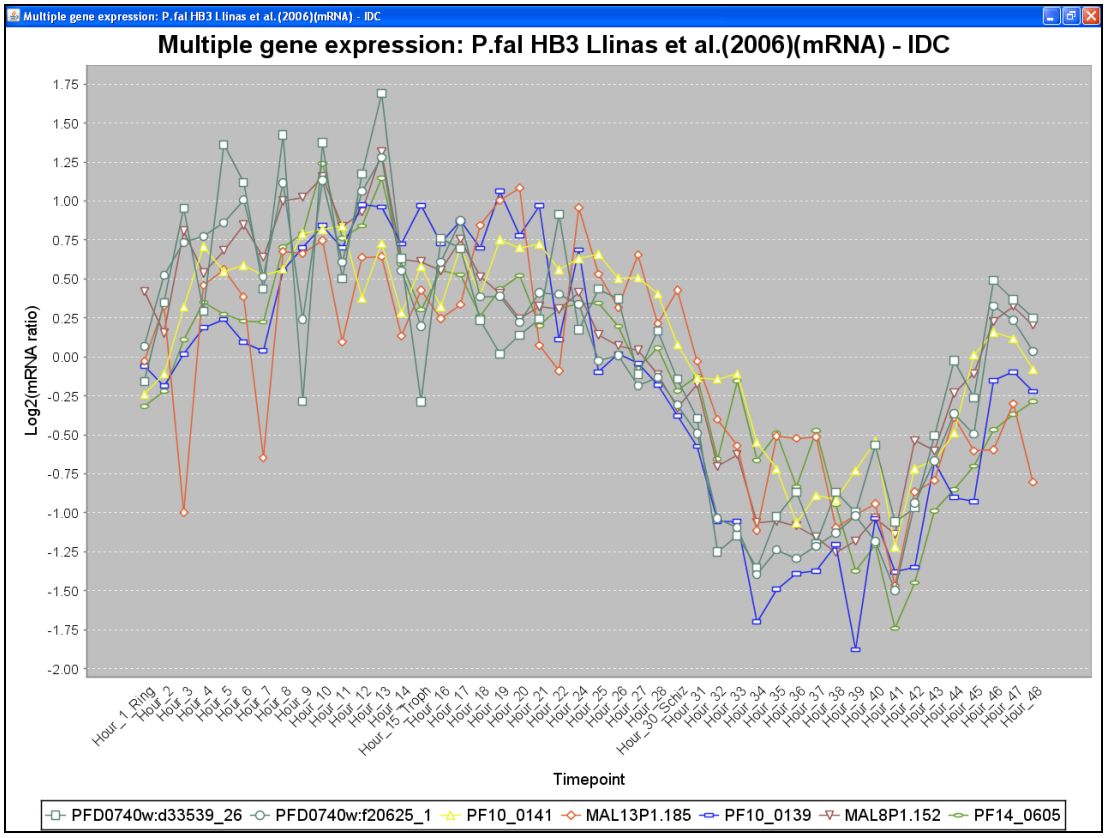


Figure 6.7. Time-series expression profiles of three CDKs [*Pf*crk-3 (PFD0740w), *Pf*mrk (PF10_0141) and *Pf*PK6 (MAL13P1.185)] and three putative cyclins [*Pf*cyc-1 (PF14_0605), PF10_0139 and MAL8P1.152] during the HB3 IDC (data from Bozdech et al. 2003).

It seems probable, though, that all these proteins play a role in regulation of RNA synthesis and/or cell growth, since they are highly expressed during the late

ring/early trophozoite stage corresponding to the period of rapid synthesis and growth (Gritzmacher and Reese 1984). As discussed above, *Pf*mrk functions as a transcription regulator through RNA polymerase II CTD phosphorylation and as such is expected to be expressed during the RNA synthesis phase.

6.1.5.5 Other observations

The only CDK/CRK in Table 6.3 not mentioned so far is *Pf*crk-1. During the IDC *Pf*crk-1 has an expression profile akin those of *Pf*PK5 and *Pf*cyc-4, peaking around the late trophozoite/early schizont stages (hours 24-44) (data not shown). In fact, *Pf*crk-1 has been demonstrated to be expressed in gametocytes (Doerig et al. 1995), so it may be only marginally expressed in the IDC (since the expression data used for this investigation do not represent absolute values (see Section 1.5.1.2) it is impossible to know the actual expression level during the IDC from this data). *Pf*crk-1 perhaps interacts with different cyclins, either from the known and predicted set or another, unidentifiable by sequence similarity, cyclin, or perhaps functions cyclin-independently at multiple stages of the life cycle.

The two cyclin-like domain-containing transcription factors, TFIIIB subunit (PF14_0469) and TFIIB (PFA0525w), generally showed little expression profile similarity to the CDKs/CRKs. There was some similarity between the expression profiles of the TFIIIB subunit with *Pf*crk-3 and between TFIIB with *Pf*mrk (data not shown). The later CRK is a known transcription regulator; therefore, it seems more likely that they function in transcription regulation rather than cell cycle control.

The Skp family protein, MAL13P1.337, a possible CDK/cyclin-associated protein (Table 6.3), shows similarity in its IDC expression profile to at least two of

the CDKs. The most strongly correlated of these is *Pf*crk-1, so if the product of MAL13P1.337 does have a CDK-regulating function, this could be its substrate.

It should also be noted that the yeast-two hybrid interaction data available for the set of known and predicted CDKs and cyclins were investigated via the MaGnET Protein-Protein Interaction Viewer, but no direct interactions were recorded (Figure 6.8). These negatives results are an indication of the unfortunate high occurrence of false negatives in this dataset. The dataset includes almost 3,000 pairwise interactions, but there are likely to be many thousands more interactions that simply were not recorded for a variety of reasons (LaCount et al. 2005).

Nonetheless, exploration of the interaction network surrounding the single CDK and two cyclin molecules that have interaction data in the LaCount et al dataset revealed that the CDK *Pf*crk-3 (PFD0740w) and the predicted cyclin encoded by MAL8P1.152 are linked in the network via a shared secondary interaction partner (PFL1385c) (Figure 6.8). Interestingly, these two proteins were observed to share a very similar expression profile during the IDC as described in Section 6.1.5.4; so this adds evidence to the theory that they might form an active complex during the IDC.

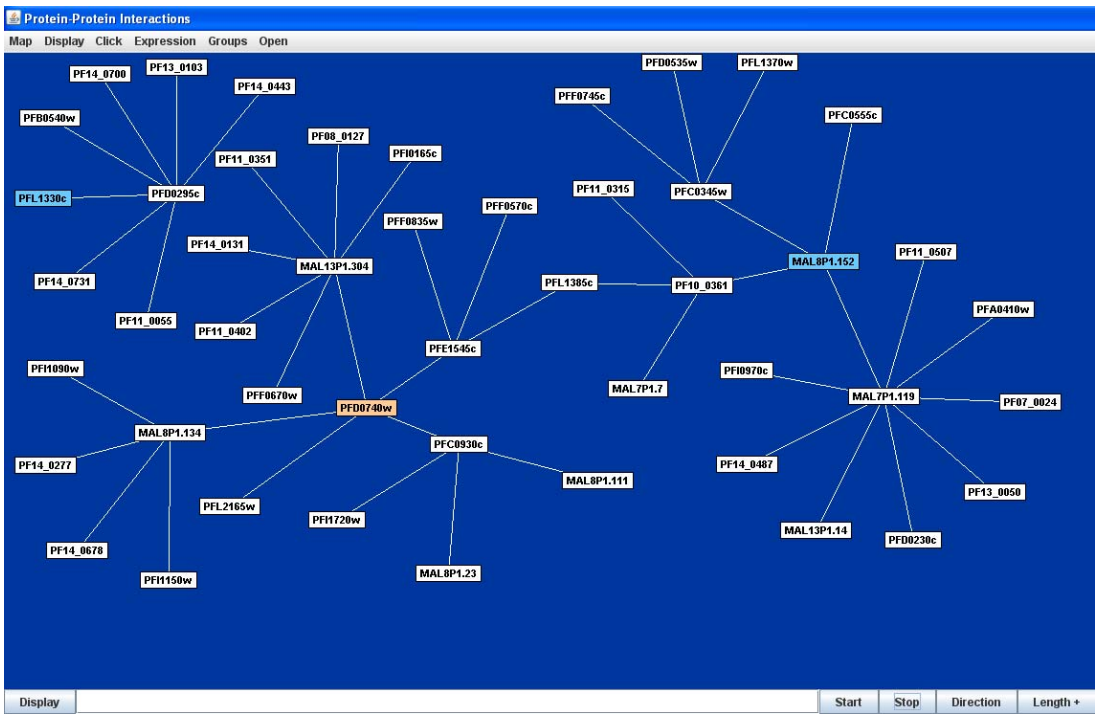


Figure 6.8. Screenshot of the Protein-Protein Interaction Viewer displaying primary and secondary interaction data for all known and predicted CDKs (orange) and cyclins (blue).

6.2 Protein-protein interaction data representing functionally-related protein clusters

Large-scale yeast two-hybrid (Y2H) screening for protein-protein interactions has proven a useful technique for enhancing understanding about topology of cellular interaction networks in a number of organisms (Koegl and Uetz 2007). Large-scale Y2H screening comes with several limitations (summarised in Section 1.5.3), but there are ways to increase confidence in interactions. These include removing non-specific interactions by filtering out 'promiscuous' proteins, removing interactions that could not be reproduced, combining multiple networks, and comparing to external data (for example, interacting proteins are more likely to be co-expressed, have related function, and be evolutionarily conserved) (Koegl and Uetz 2007).

Since there is only one large-scale Y2H dataset currently publicly available for *Plasmodium*, there is no opportunity for combining results from multiple experiments to strengthen the interaction information. However, the other techniques above can be employed to discover interactions within the dataset that are likely to be true results. The next section will describe examples where this has been successful for the analysis of small interaction clusters, ultimately leading to postulations about the functions of some uncharacterised proteins in the cluster.

6.2.1 Predicting function of hypothetical proteins in a cluster of interacting proteins with characterised function

Exploration of the *P. falciparum* protein-protein interaction data (LaCount et al. 2005) with the MaGnET Protein-Protein Interaction Viewer revealed an apparent abundance of data for proteins associated with cytoplasmic and nuclear substructures performing core metabolic functions, such as the ribosome, spliceosome, proteasome and nucleosome (data not shown). This is not surprising given that intracellular, non-membrane spanning proteins are more likely than membrane spanning or secreted proteins to be identified in yeast two-hybrid and similar experiments (Koegl and Uetz 2007).

Since many of the proteins found in this interaction subset have been experimentally characterised or have strong similarity to known proteins in other organisms, hypotheses may be generated about the potential function of hypothetical proteins occurring within the network. Examples of how such hypotheses can be generated are discussed in the following sections and evidence is provided backing up the hypotheses.

6.2.1.1 Identification of a novel putative intracellular protein hypothesised to regulate a number of processes including protein metabolism and gene expression

The gene PFI1715w encodes a hypothetical protein with no detectable similarity to known proteins. This protein was found to directly interact with 27 proteins in yeast two hybrid experiments (Figure 6.9) (LaCount et al. 2005). Of these, 16 interactions were recorded more than once and six of these were recorded more than five times. When interactions are repeatable it increases confidence that the interaction is real and not a false positive result. This set of repeatable interactions can provide useful information about the probable protein function of PFI1715w, since proteins that interact are more likely to have related function (Koegl and Uetz 2007).

The PFI1715w protein's repeatable interaction partners include two DNA-binding proteins: a bromodomain protein (PFL0635c) – bromodomain-containing proteins are involved in regulating chromatin structure and hence gene expression (Marmorstein and Berger 2001) – and a protein with similarity to CCAAT-box DNA-binding protein subunit B (MAL13P1.21) – a transcription factor. The former interaction was recorded 18 times and the latter 7 times, so there is good evidence for these being true interactions. Interactions were recorded with several other DNA-binding proteins, although in most cases these were not repeatable, so must be treated with caution. The putative binding partners include three helicases [PF10_0232, PF11_0053 (both recorded once) and PFF1185w (recorded 3 times)]. An interaction was also recorded once with a second CCAAT-box DNA-binding protein subunit B (PF11_0477).

The PFI1715w product also appears to be linked to a number of metabolic pathways. Its interaction data include several interactions with components of protein metabolic pathways; however it should be noted that most of the interactions were either not repeatable or repeated only once or twice. They include: two ribosomal subunit proteins (PFE0350c and PFF0885w), ribosome biogenesis regulatory protein (PF11_0259), two splicing factors (PFI1115c and PFE0865c), a putative PRP4 (pre-mRNA processing factor 4) kinase (PF11_0156) and a proteasome subunit (PF07_0112). Of this list, one of the interactions was repeatable 12 times: that with the splicing factor encoded by gene PFI1115c; this interaction is likely to be a true positive.

In addition, PFI1715w may be linked to purine metabolism via a possible interaction with a putative allantoicase protein (PF07_0120) and to lipid metabolism via a possible interaction with a phospholipase (PFB0870w). Both of these interactions were recorded only once though so they may be false positives.

An interaction was recorded 10 times between the PFI1715w protein and the protein encoded by PFI1680w – a probable FAS-associated factor (FAF). FAF is a multi-functional protein: it is a member of the apoptosis-inducing signalling complex (Ryu et al. 2003); it is thought to play a role in regulation of the ubiquitin-proteasome pathway (Song et al. 2005); it can also inhibit heat shock protein 70 (HSP70) chaperone activity (Kim et al. 2005). Therefore, it is plausible that PFI1715w could influence several intracellular signalling pathways.

Another interaction partner that was recorded many times was the 14-3-3 protein (MAL8P1.69), a mediator for signalling pathways and regulation of cell cycle control through protein binding (Al-Khedery et al. 1999). There is also a small

amount of further evidence linking PFI1715w to cell cycle control through a possible interaction (recorded twice) with a component of the gamma-tubulin complex of the spindle pole (PF14_0414).

Overall, it seems possible that PFI1715w may be important for cross-talk between factors involved in intracellular signalling pathways controlling gene expression and protein metabolism pathways; for example, by interacting with transcription factors, helicases and splicing factors.

The MaGnET gene fact sheet for PFI1715w showed that its predicted protein sequence does not have any recognisable signal sequences (data not shown). Therefore, it is unlikely to be exported out of the parasite. It may be located in the nucleus, as it appears to interact with DNA-binding proteins. To ascertain if there was further evidence for its nuclear location the predicted protein sequence was submitted to the ScanProsite protein domain and motif detection server (de Castro et al. 2006). No bipartite nuclear localisation signals were predicted, which does not support a nuclear location for this protein. However, the protein could feasibly get into the nucleus by forming a complex with other proteins that do have nuclear localisation signals. The PFI17175w protein may be located in the cytoplasm until it binds to protein partners that cause it to translocate to the nucleus.

Since protein-protein interaction data generated by yeast two-hybrid experiments is known to contain a relatively high number of false positives, it should always be treated with care. That MaGnET provides data about the number of independent searches and number of times each interaction was observed on the gene fact sheets is very useful for assessing the quality of the evidence about an interaction. Some of the interactions observed for PFI1715w may turn out to be false

positives; indeed, there is also weak evidence for an interaction with *Pf*EMP1 (PFF1580c) and reticulocyte binding-like protein 3 (PFL2520w) which look like outliers compared to its other interaction partners.

One of the advantages of MaGnET is that it brings together multiple data-types in one place, so allows the exploration of all lines of evidence during hypothesis generation. Exploration of the expression data surrounding the PFI1715w gene reveals that it is expressed mainly during the schizont stage of the IDC (Figure 6.9). Expression during the schizont stage would concur with a potential role in regulation of processes linked to cell cycle, gene expression and protein metabolism, during which time the parasite is undergoing rapid rounds of mitosis and forming new merozoites ready for release at the end of the schizont stage.

Figure 6.9 demonstrates that the other members of the interaction cluster generally follow a similar expression profile to PFI1715w in the IDC. The majority are also expressed during the schizont stage. A glaring exception is PFF1580c encoding *Pf*EMP1, which is not surprising since during the IDC *Pf*EMP1 is expressed during the late ring/early trophozoite stage. Another exception is the ribosome biogenesis regulatory protein (PF11_0259), because peak expression of this protein occurs earlier in the IDC, presumably during the time-frame of ribosome synthesis. The fact that expression of PFI1715w does not overlap with that of PFF1580c and PF11_0259 in the IDC indicates that these interactions could be false positives, since they are unlikely to meet under biological conditions (notwithstanding the possibility that they may be co-expressed and interact during other phases of the life-cycle).

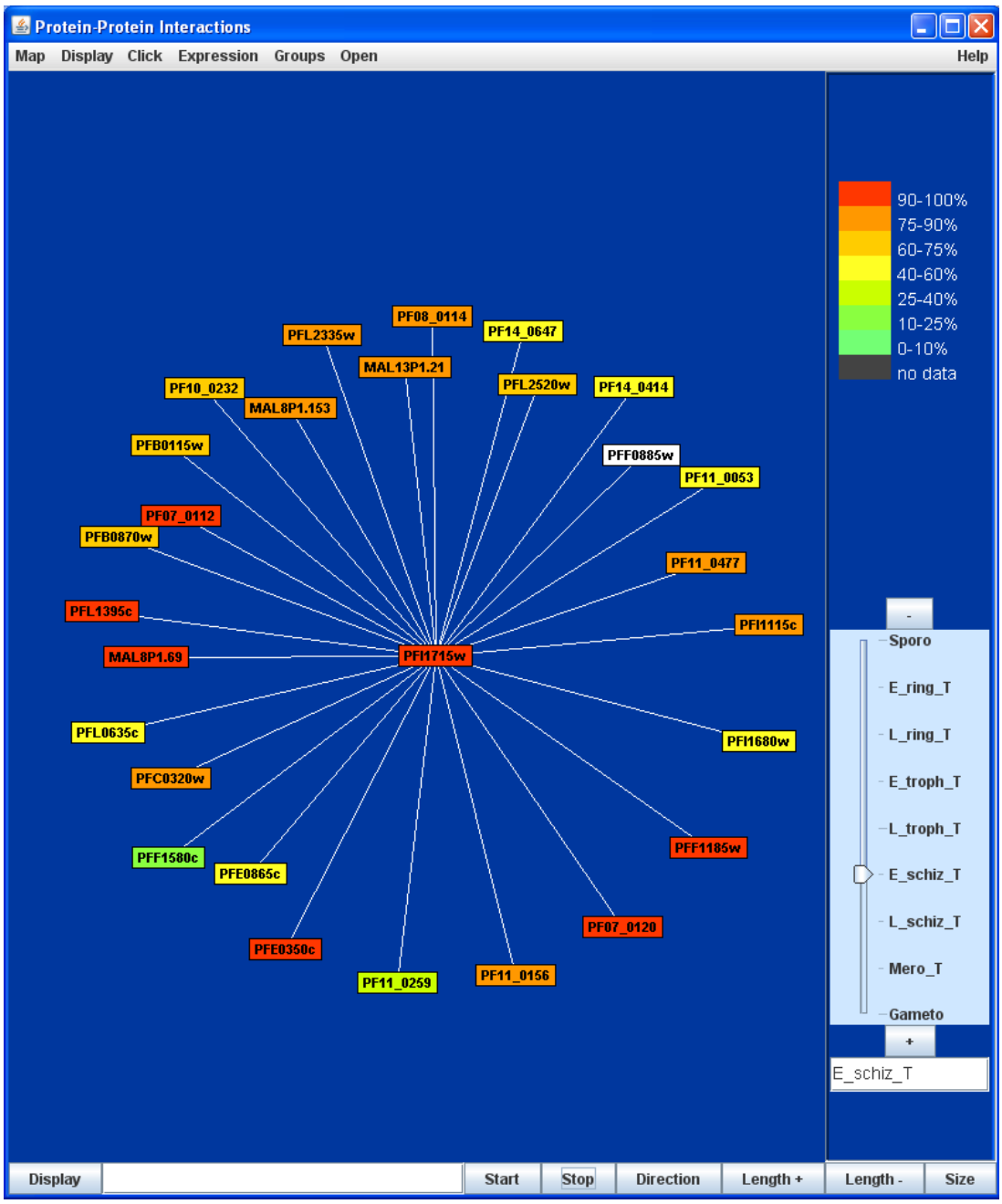


Figure 6.9. Primary interactions of the PFI1715w protein with transcription levels at the early schizont stage overlaid (interaction data from LaCount et al. 2005; transcription data from Le Roch et al. 2003).

To assess the validity of the observations made about the PFI1517w protein, further analysis of the functional enrichment of its interaction partners was

undertaken. The CLENCH2.0 program (Shah and Fedoroff 2004) was used to calculate the GO categories that are significantly enriched in this set of proteins. The interaction partners of PFI1715w are enriched for categories related to regulation of metabolic processes, RNA metabolism, chromatin remodelling, regulation of transcription, RNA splicing and nuclear and ribosomal localisation (Table 6.4).

Term	Aspect	P-value
Regulation of biological process	Biological process	0.021
Regulation of cellular metabolic process	Biological process	0.010
Organelle organisation and biogenesis	Biological process	0.001
Chromatin remodelling	Biological process	0.000
Macromolecule metabolic process	Biological process	0.046
Biopolymer metabolic process	Biological process	0.021
RNA metabolic process	Biological process	0.010
Regulation of transcription, DNA-dependent	Biological process	0.003
RNA splicing	Biological process	0.016
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	Biological process	0.033
Nucleic acid binding	Molecular function	0.046
DNA-binding	Molecular function	0.010
Helicase activity	Molecular function	0.011
Ribonucleoprotein complex	Cellular component	0.020
Large ribosomal subunit	Cellular component	0.014
Spliceosome	Cellular component	0.007
Nucleus	Cellular component	0.000
Chromatin remodelling complex	Cellular component	0.001

Table 6.4. A representative selection of the enriched GO categories for the group of proteins involved in primary interactions with the protein encoded by PFI1715w. At a confidence level of 95%, a p-value of below 0.05 indicates that a category is significantly enriched in this set compared to all *P. falciparum* proteins. (Full analysis results are included on the accompanying CD.)

6.2.1.2 Identification of a novel putative nuclear protein hypothesised to regulate protein metabolism and chromatin modification

The final interaction of the PFI1715w protein that was recorded several times was with a hypothetical protein encoded by the gene MAL8P1.153. Like PFI1715w,

this protein was reported to interact with many nuclear proteins, several of which were repeatable (LaCount et al. 2005). Some of the interacting proteins have function annotated, so the interaction data for MAL8P1.153 was explored to see whether any clues could be obtained about its possible function.

A self-interaction of the MAL8P1.153 protein was recorded 70 times in the yeast-two hybrid data. Therefore, MAL8P1.153 proteins certainly form a homo-subunit complex, which is probably required for its function.

From the protein-protein interaction data there is a good amount of evidence indicating that MAL8P1.153 may function in regulating protein and nucleic acid metabolism. MAL8P1.153 was linked to protein catabolism through an interaction with a probable ubiquitin carboxyl-terminal hydrolase family 2 protein (PFI0225w), which was recorded 11 times. There was also an interaction with a putative ubiquitin-protein ligase (MAL8P1.23), but that was recorded only once, so it could be a false positive. Additionally, an interaction between the MAL8P1.153 protein and a protein involved in mRNA degradation – CAF1 family ribonuclease (MAL8P1.104) was observed 5 times.

There is limited evidence for a link to factors involved in protein synthesis, but as the interactions were not repeatable this should be treated with caution. The interaction partners included a homologue of human HSPC025 (PFF0590c), otherwise known as eukaryotic translation initiation factor 3, subunit E interacting protein (EIF3EIP), and a Sec63 homolog (PF13_0102) – a member of a complex that mediates newly synthesised protein transport into the endoplasmic reticulum.

The MAL8P1.153 protein may also interact with other transmembrane transporters, as an interaction with a sulphate transporter (PF14_0679) was recorded

four times and an interaction with a potassium channel (PFL1315w) was recorded twice.

MAL8P1.153 may link to chromatin modification and DNA synthesis pathways through interactions with histone acetyltransferase Gcn5 (PF08_0034) – required for chromatin remodelling and transcription activation (Fan et al. 2004) (recorded 5 times), a helicase (PF10_0232) and DNA polymerase epsilon, catalytic subunit A (PFF1470c) (both recorded just once).

These observations suggest that one or more MAL8P1.153 subunits combine to form a functional complex, with or without other factors. The active complex may be involved in regulation of protein metabolism and could be involved in cross-talk between these pathways and those controlling gene transcription and DNA synthesis. Therefore, it may require shuttling between the cytoplasm and the nucleus to carry out its function.

Examination of the expression profile of MAL8P1.153 (Figure 6.10) reveals that the gene is transcribed at several distinct stages of the life cycle and that its expression is quickly switched on and off, indicating that its transcription is tightly controlled. MAL8P1.153 is expressed during the stages of the life cycle when the parasite is undergoing phases of replication and differentiation (sporozoite, schizont and gametocyte), which indicates that this protein might be important during these processes. If MAL8P1.153 is expressed at several stages of the life cycle, it could conceivably interact with proteins that are present at different times, as is suggested by Figure 6.11, which shows that different interacting partners are expressed at different stages.

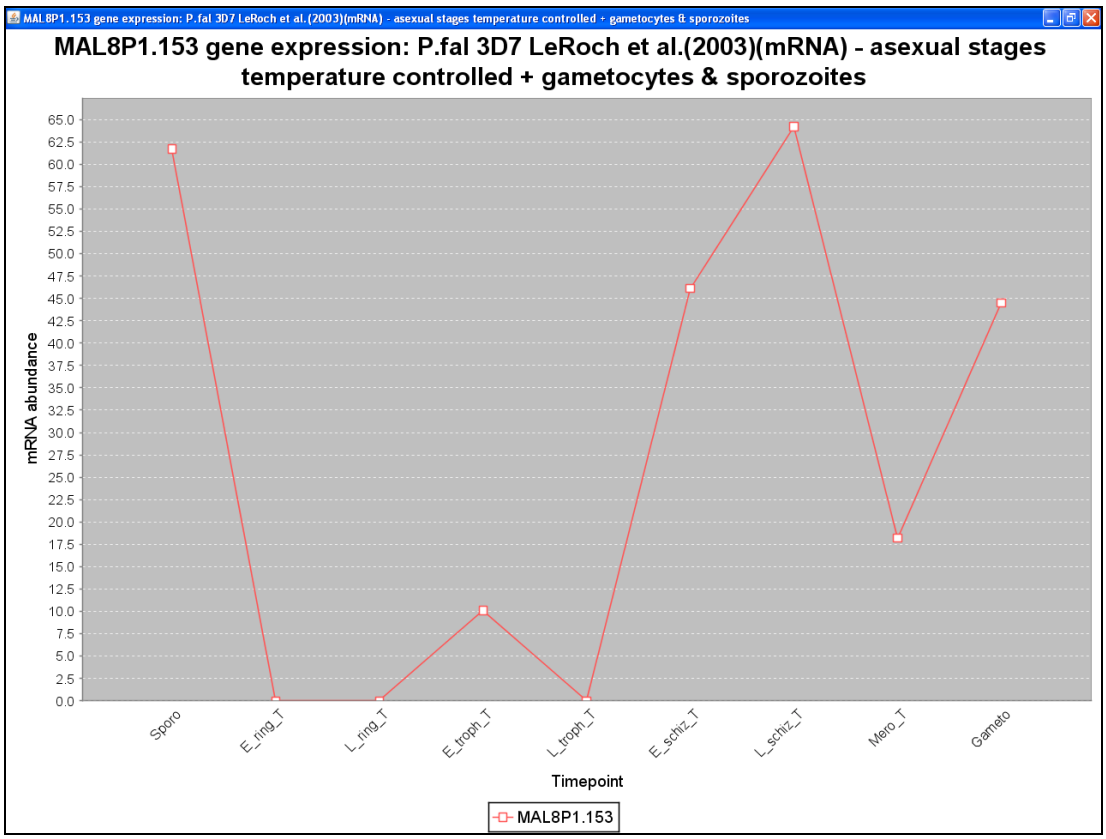


Figure 6.10. Time-series expression profile of the MAL8P1.153 gene (data from Le Roch et al. 2003). Transcription peaks during several stages of the life cycle, including the sporozoite, schizont and gametocyte stages.

To gather further evidence about the function of MAL8P1.153, its predicted protein sequence was submitted to the ScanProsite protein domain and motif detection server (de Castro et al. 2006). The results revealed that the MAL8P1.153 protein has two bipartite nuclear localisation signals, indicating that it is likely to be localised in the nucleus. There were no regions enriched for positively charged residues, indicating that the protein would be unlikely to bind DNA (negatively charged phosphate groups of DNA bind suitably spaced positively charged protein side chains). This suggests that any role it may play in regulating DNA synthesis or gene expression must be asserted through interactions with other proteins.

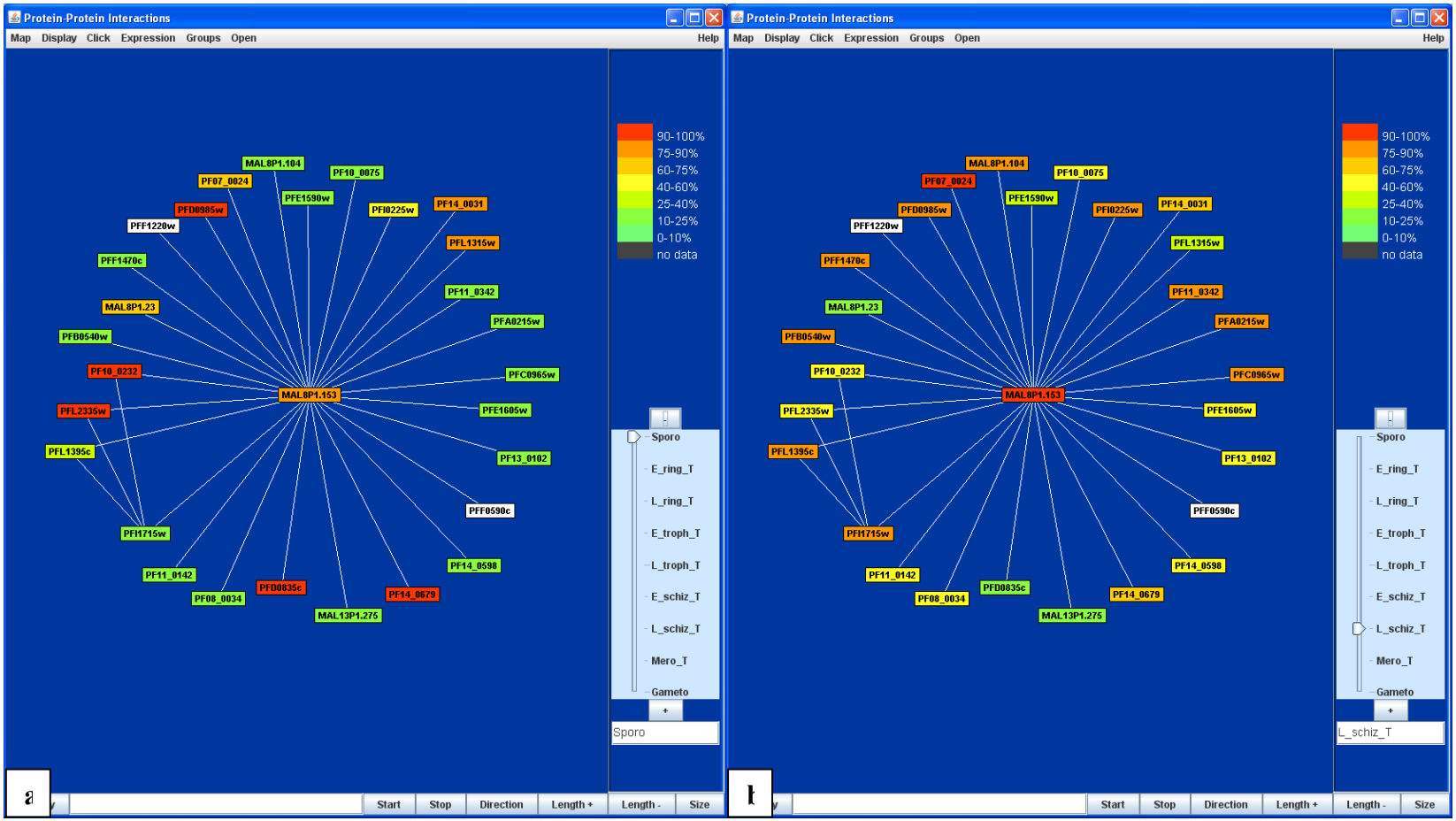


Figure 6.11. Protein-protein interactions of the protein encoded by gene MAL13P1.153. Proteins are coloured according to their transcription level at (a) the sporozoite stage and (b) the late schizont stage (data from Le Roch et al. 2003).

6.2.1.3 Identification of a putative novel DNA-binding protein

The gene PF10_0232 encodes a protein known as chromodomain-helicase-DNA-binding protein 1 (CHD1) homolog, which has a variety of roles depending on its interaction partners, including transcription activation and repression and mRNA splicing [reviewed in (Hall and Georgel 2007)]. As shown in Figure 6.11, CHD1 is one of three protein interaction partners shared between the two hypothetical proteins encoded by PFI1715w and MAL8P1.153 (described above). The other two shared interaction partners are hypothetical proteins encoded by PFL2335w and PFL1395c. As Figure 6.11 reveals the expression of CHD1 (PF10_0232) and PFL2335w appear closely coupled at two life cycle stages. Inspection of their expression profiles reveals that they are actually highly similar across sporozoites, the IDC and gametocytes (Figure 6.12). The evidence for an association between CHD1 and PFL2335w is further strengthened by the fact that they also have a third interaction partner (a hypothetical protein encoded by PF14_0499) in common between them. A direct interaction between the CHD1 and PFL2335w-encoded proteins has not been recorded; however, proteins with similar functions are more likely to share interaction partners (Koegl and Uetz 2007). Since these two proteins have a similar expression profile, and share several interaction partners, it can be hypothesised that they share a related function. The PFL2335w protein is probably a DNA-binding protein functioning in transcription regulation, probably in complex with other proteins.

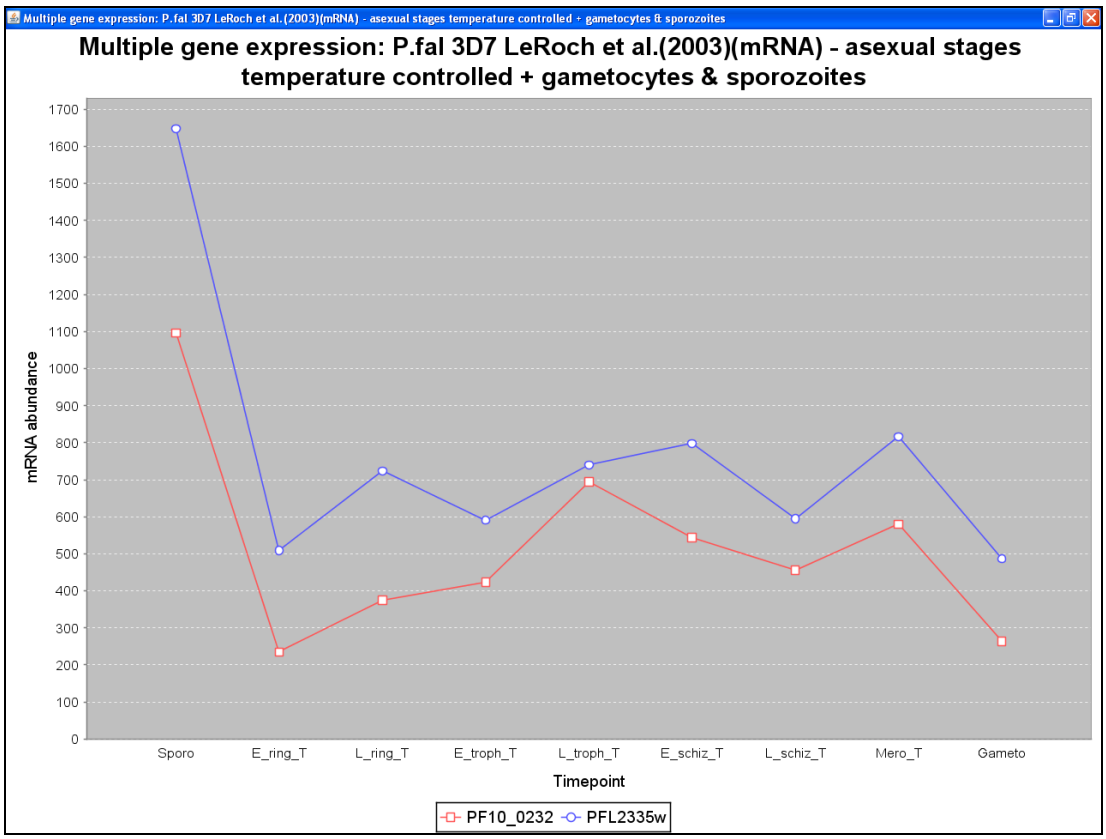


Figure 6.12. Expression profiles of the genes encoding CHD1 (PF10_0232) and a hypothetical protein (PFL2335w) (data from Le Roch et al. 2003).

To examine further evidence for this hypothesis the predicted cross-organism orthologs for the PFL2335w gene were accessed using a link from its MaGnET gene fact sheet to its entry in OrthoMCL-DB (Chen et al. 2006a). The predicted orthologs for this gene include a DNA-binding domain containing protein from *Oryza sativa*.

The protein sequence of PFL2335w was submitted to the ScanProsite motif detection server (de Castro et al. 2006). The results showed that there are two lysine-rich regions within the sequence, confirming that the protein has the necessary positively charged residues to support a potential electrostatic interaction with DNA (data not shown). Other predicted sequence motifs included bipartite nuclear localisation signals and an EF-hand calcium-binding domain.

Taken together, this evidence suggests that the protein encoded by PFL2335w is likely to be a novel DNA-binding protein, possibly working in concert with CHD1, regulating gene expression within the sporozoite and intraerythrocytic stages of the parasite's life cycle.

6.3 Exploring characteristics of species-specific gene families with high numbers of pseudogenes

There are two mechanisms by which pseudogenes can be generated in eukaryotes: gene duplication ('duplicated pseudogenes') and subsequent disabling of one gene copy, and reverse transcription of mRNA randomly inserting the sequence into genomic DNA ('processed pseudogenes'). Both have distinct characteristics: duplicated pseudogenes are usually disabled by stop codons or frame-shifts; processed pseudogenes tend to lack introns and may contain other artefacts of transcription, such as polyadenine tails (Harrison and Gerstein 2002).

Processed pseudogenes are more prevalent in gene families that are highly expressed, due to the large amount of mRNA that is available for random reverse transcription and insertion events (Harrison and Gerstein 2002). Duplicated pseudogenes, however, tend to occur more frequently in organism-specific families that are linked to environmental response functions (Harrison and Gerstein 2002).

There remains the possibility that pseudogenes can be resurrected as new proteins after undergoing a period of random drift without selection. Reservoirs of pseudogenes will therefore increase the potential sampling space for proteome evolution (Harrison and Gerstein 2002).

It was decided to investigate the relationships between pseudogenes and gene families within *P. falciparum* using tools available within MaGnET.

A well-known multi-copy gene family that is unique to *P. falciparum* are the *var* genes, encoding *Pf*EMP1, a protein expressed on the surfaces of infected erythrocytes (IE) and mediating cytoadhesion. *Var* genes undergo antigenic switching, so that usually one family member is dominantly expressed per parasite generation (Peters et al. 2002). The red line in Figure 6.13 represents the dominantly expressed full-length *var* gene in one particular microarray experiment.

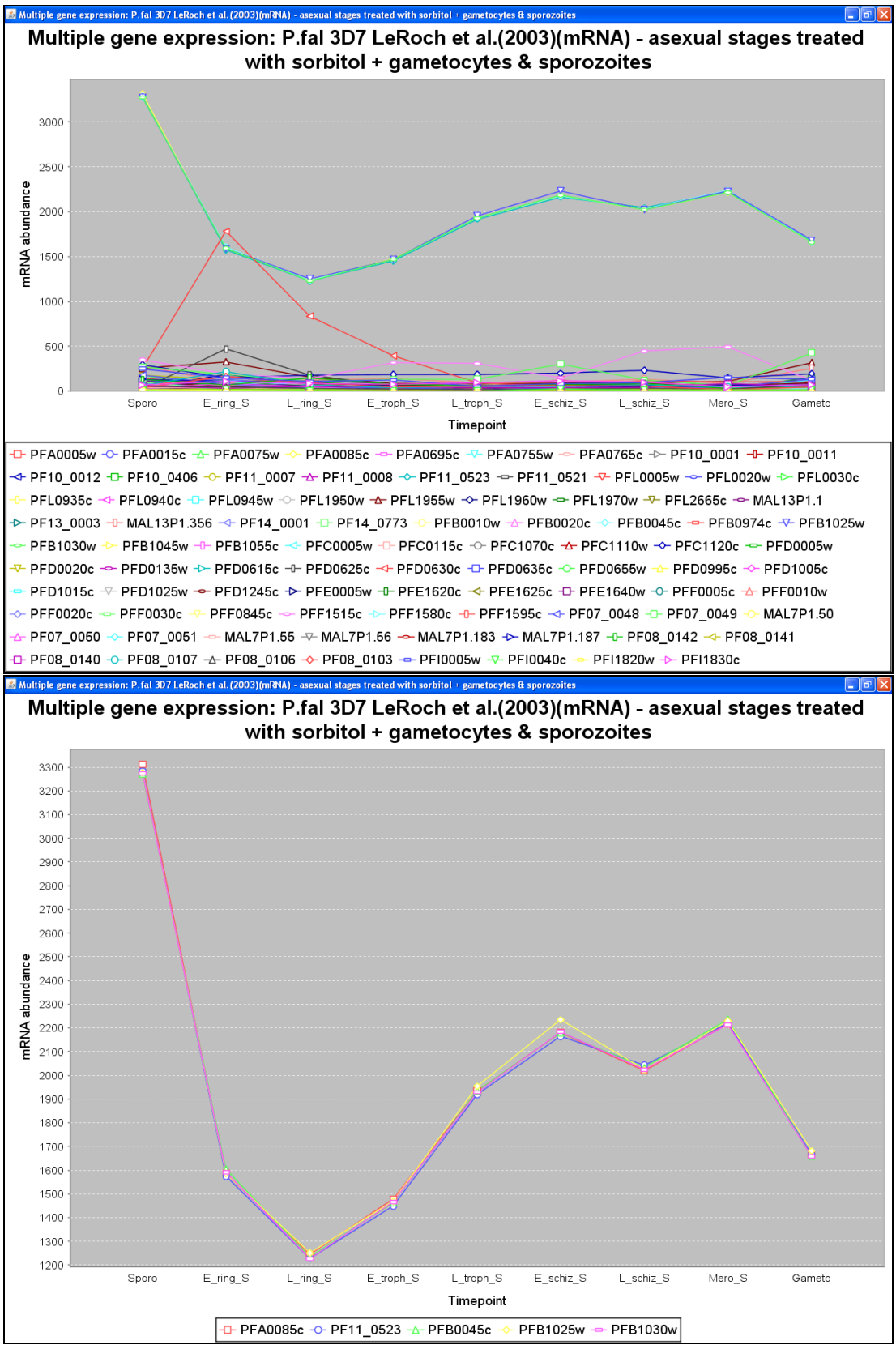


Figure 6.13. Graph showing the expression profiles of all predicted *var* genes (including pseudogenes) encoded by the *P. falciparum* 3D7 genome that had expression data recorded in a study

by Le Roch et al. 2003 (top panel). The red line indicates the dominantly expressed *var* gene. The graph also shows apparent constitutive, high expression of a group of five probes mapping to *var* pseudogene sequences, probably caused by cross-hybridisation to the probes by a set of similar transcripts. For clarity, the bottom panel shows a close-up view of these five profiles.

The surprising aspect of *var* gene expression revealed by Figure 6.13 is the apparent constitutive high expression of five other *var* genes, following a completely unrelated expression profile to that normally seen for *var* genes (*var* gene transcription is normally switched on in ring stage parasites). These five highly expressed genes are in fact all annotated as pseudogenes, degenerate or truncated *var* genes. Each gene is also only represented by one probe on the microarray used in this study (Le Roch et al. 2003). Therefore, the constant, high signal from these probes is surely due to cross-hybridisation of a set of similar transcripts. A similar phenomenon involving the same five probes is also observed with data from the gametocyte stages of 3D7 and NF54 strain parasites (data not shown) (Young et al. 2005). This indicates that many similar *var* genes (and perhaps pseudogenes) are being transcribed but not translated in several *P. falciparum* life cycle stages. The biological advantage of this to the parasite is unclear, but may be linked to the large number of truncated *var* pseudogenes in the genome, which could have arisen out of reverse transcription and insertion of spliced sequences back into the genome. One effect of this could be to ensure an expansive available reservoir of sequences from which to generate possible new proteins; bestowing an advantage for immune evasion.

To investigate whether there were any genome-wide patterns for gene families with pseudogenes being more abundantly expressed, a custom Java script

was developed to calculate overall family expression levels and compare them to the number of pseudogenes. Expression level data for each family member was averaged over a time-series, family members' expression levels were totalled, and then the totals from two experiments (Le Roch et al. 2003; Young et al. 2005) were combined to give an overall representation of mRNA abundance for each family. *P. falciparum* gene families were identified as being the paralogous genes within individual clusters of homologues as calculated by the OrthoMCL program (data stored in the MaGnET database – see Chapter 3 for more details) (Li et al. 2003).

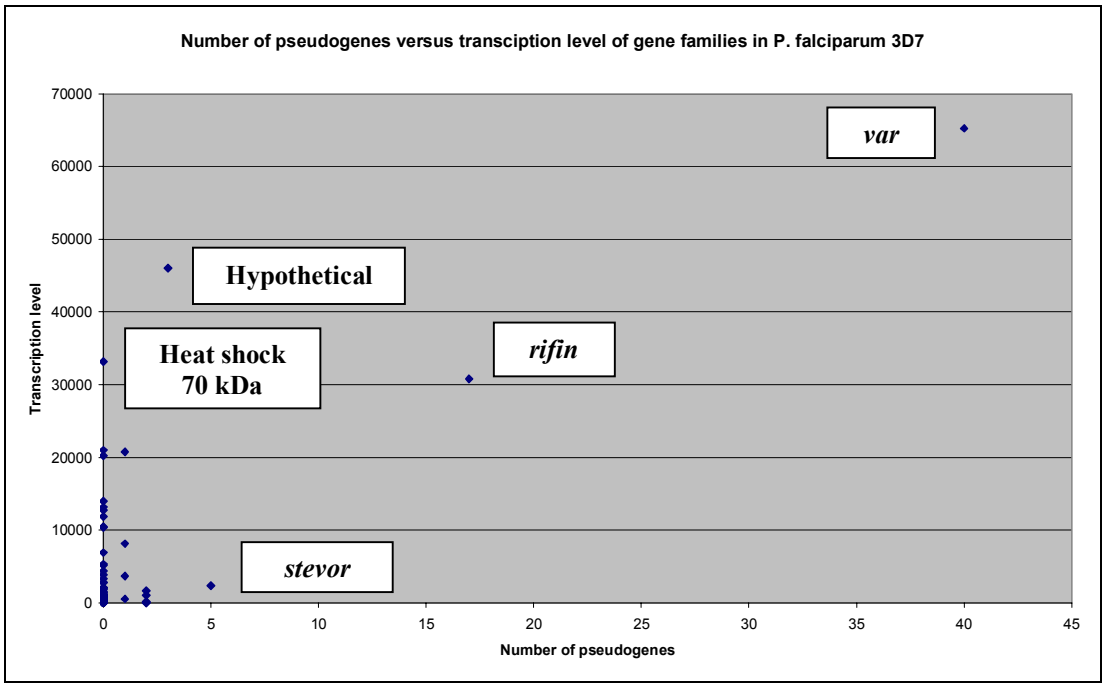


Figure 6.14. Graph showing overall expression level of *P. falciparum* 3D7 gene families plotted against number of pseudogenes in the family. Expression level was calculated by summing average expression levels of family members across a time-series experiment and combining the totals from four time-series experiments that used the same array (Le Roch et al. 2003; Young et al. 2005).

Figure 6.14 presents a graph of the results. The graph shows that on average, there are fewer than 5 pseudogenes per family in the *P. falciparum* genome, and just two families contain the majority of all pseudogenes. The family with the second highest number of pseudogenes after *var* is *rifin*, another species-specific antigenic family, and this family is the fourth most abundantly expressed. However, the family with the third greatest number of pseudogenes, *stevor*, a species-specific family related to the *rifins*, appears to buck the trend. *Stevor* is known to be expressed during the life cycle stages sampled (namely those occurring in the human host) (McRobert et al. 2004), but little *stevor* expression has been detected with microarray studies (Le Roch et al. 2003; Llinas et al. 2006). Therefore, this result may just reflect that the current arrays do not accurately represent real *stevor* genes and fail to capture their true expression level. The family with the fourth highest number of pseudogenes, a family of hypothetical protein encoding genes, also follows the trend by having the second highest overall expression level.

Overall, there seems to be a relationship between large, species-specific families having high transcript abundance, and the number of pseudogenes per family. It would be interesting to investigate further the nature and characteristics of the pseudogenes to determine what proportion of them arose by gene duplication or processing of transcripts and how rapid the turn-over of genes is in *P. falciparum*. Comparison of pseudogene complement between strains and species will be important for understanding their evolution.

6.4 Identifying cases of misannotation

When genomes are annotated, names are often assigned to genes on the basis of similarity to other proteins. Often, the similarity may only encompass part of the protein; for example, a single domain within a multi-domain protein. Therefore, it can be dangerous to assign a function when the proteins do not have similar global arrangements of domains because the new sequence may be missing vital functional regions. Genome annotation often involves automated assignment of probable protein domains using databases of known domain and protein families, such as InterPro (Mulder et al. 2007). Gene classification systems such as Gene Ontology (Ashburner et al. 2000) have evolved to provide a method of annotating gene function within a controlled vocabulary along with a means to document the source of annotation. Logically, the next step would be to combine automated domain annotation with a description of the protein's function using ontologies.

For several years following release of the sequence of the *P. falciparum* 3D7 genome (Gardner et al. 2002), the GO annotation remained relatively sparse due to the high number of hypothetical genes. Recently, the GO annotation has been bolstered by introduction of automatic assignment of GO terms based on the presence of a particular domain predicted by InterPro. During the course of this work many incidences of questionable GO annotation have come to light, which have arisen through the inappropriate assignment of terms when a particular domain has been predicted in a protein sequence. MaGnET presents the annotation data in a way that makes it easy to find cases of probable misannotation. Other online genome resources, such as PlasmoDB and GeneDB, that include similar data, do not always

clearly provide the evidence for functional assignments in a way that makes it easy for the user to trace the origin of GO terms. An example is discussed below.

6.4.1 Example: a misannotated potassium channel

The gene PFL1315w encodes a potassium channel (*Pf*k1) (Waller et al. 2008). Figure 6.15 shows the GO annotation and InterPro predicted protein domains for PFL1315w displayed in the gene fact sheet. The InterPro annotation includes several hits to potassium channels from various sources, most of which are in agreement over the location of a characteristic sequence motif of potassium channels within the protein sequence. The InterPro annotation also includes a hit to a zinc protease motif towards the C-terminal. The zinc protease motif hit led to automatic assignment of three GO terms to the gene ("metallopeptidase activity", "zinc ion binding" and "proteolysis") (Figure 6.15). Two of these three GO terms are misplaced, since some potassium channels do require zinc binding, but they do not function as metallopeptidases. This highlights an error that can be caused by computationally assigning function to a gene based on a hit to a single, short motif within a large sequence.

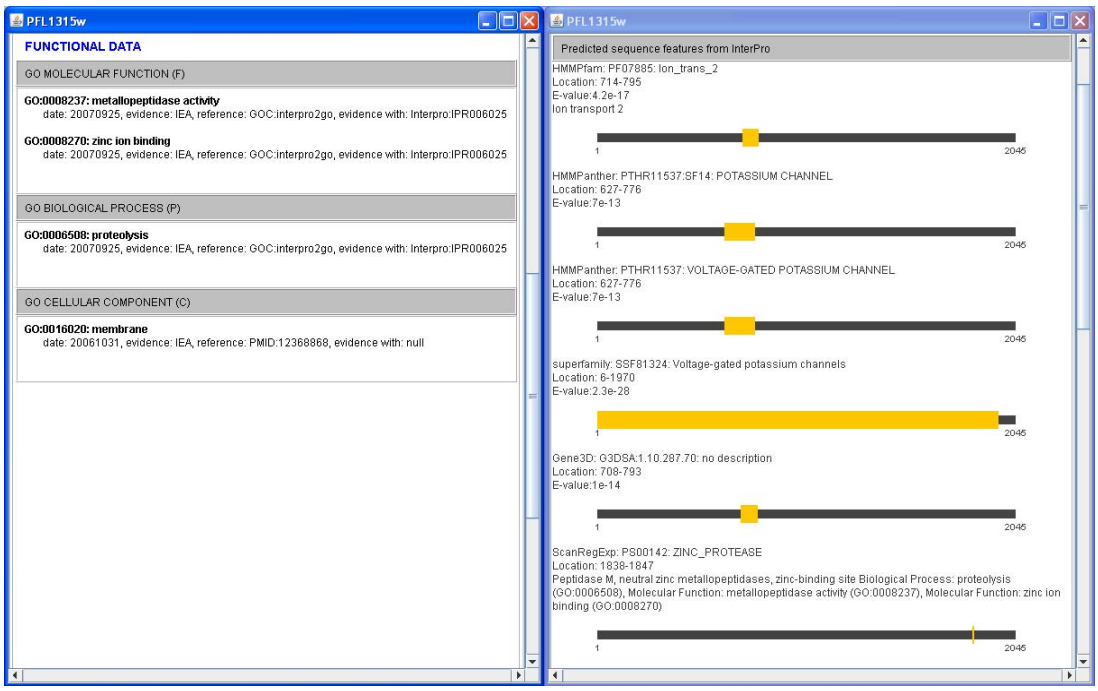


Figure 6.15. Left panel: the GO annotation assigned to *Pf*k1. Right panel: the InterPro predicted protein domain and sequence features for *Pf*k1.

The GO annotation includes only one other term associated with the correct function of the protein ("membrane"), which does not adequately describe its potassium channel function since it only indicates general membrane localisation. It is unclear why the multiple InterPro hits to potassium channels did not lead the annotation software to deduce that GO terms associated with potassium channel function would be appropriate. That the software would automatically assign GO terms based on zinc protease function from a single hit but not for potassium channel function where several methods were in consensus, would seem like a gross oversight of the method. A more effective and accurate system should take into account the actual combination of domains and motifs and their prediction scores in order to assign function.

In summary, Figure 6.15 demonstrates that MaGnET provides essential data about the source and confidence scores of the predicted InterPro and GO annotation included in the MaGnET database. MaGnET provides the necessary information to users in order that they can decide how much weight to place on individual annotations and how reliable they are likely to be. In this way, users will easily be able to notice cases of misannotation, such as that described above, but also cases of more ambiguous annotation that can be further investigated using other tools.

6.5 Discussion

The mini-studies described in this chapter serve to demonstrate how exploration of functional genomic data using MaGnET can lead to new hypotheses about *P. falciparum* gene function. Table 6.5 summarises all the hypotheses and observations that have come out of the analyses using MaGnET. The hypotheses range from predictions about the putative function of hypothetical proteins to prediction of protein complex subunits to observations about properties of species-specific gene families.

Genes	**Hypothesis**
PF10_0141 (*Pf*mrk), PF14_0605 (P*f*cyc-1) and PFE0610c (*Pf*MAT1)	The RNA polymerase II CTD phosphorylation complex components are differentially co-expressed in the trophozoite stage (timing of abundant mRNA synthesis), so they are involved in regulating transcription, and hence cell growth, during this life cycle stage.
MAL13P1.279 (*Pf*PK5) and PF13_0022 (*Pf*cyc-4)	During the erythrocytic stages, the cyclin Pfcyc-4 partners with the CDK *Pf*PK5.
MAL13P1.279 (*Pf*PK5)	*Pf*PK5 forms complexes with different cyclins at different stages of the life cycle, with distinct functions.
MAL13P1.279 (*Pf*PK5) and PFL1330c (*Pf*cyc-2)	*Pf*PK5 and *Pf*cyc-2 may form a complex in gametocyte stages.

MAL13P1.279 (*Pf*PK5) and PFE0920c (*Pf*cyc-3)	*Pf*PK5 and *Pf*cyc-3 may form a complex at other stages of the life cycle, e.g. in the mosquito.
PFC0755c (*Pf*crk-4), PFF0750w (*Pf*crk-5), MAL13P1.196, PFL1330c (*Pf*cyc-2), PFF0280c and MAL13P1.131	3 CDKs and 3 cyclins are differentially co-expressed in schizonts and are involved in regulating processes linked to cell cycle.
	Waves of tightly regulated CDK/Cyclin expression provide a means of controlling cell cycle progression in schizonts.
PFD0740w (*Pf*crk-3), PF10_0141 (*Pf*mrk) and MAL13P1.185 (*Pf*PK6), PF10_0139, MAL8P1.152 and PF14_0605 (*Pf*cyc-1)	3 CDKs and 3 cyclins are differentially co-expressed during the ring/trophozoite stages and are involved in regulating transcription and cell growth.
MAL13P1.185 (*Pf*PK6)	*Pf*PK6 activity may be enhanced by binding a cyclin and it may bind a product of either PF10_0139, MAL8P1.152 or PF14_0605 (*Pf*cyc-1) during ring/trophozoite stages.
MAL8P1.152 and PFD0740w (*Pf*crk-3)	The CDK Pfcrk-3 and the putative cyclin encoded by MAL8P1.152 may form a complex during erythrocytic stages.
PFD0865c (*Pf*crk-1) and MAL13P1.337 (Skp1)	The putative Skp1 protein possibly binds to and activates *Pf*crk-1. *Pf*crk-1 may not require a cyclin to be active, or may bind a novel cyclin-like protein.
PFI1715w	This hypothetical protein-coding gene may function in mediating cross-talk between several fundamental intracellular pathways; in particular there is evidence for its function in regulating gene expression and protein metabolism pathways during the IDC.
PFI1715w	The PFI1715w-encoded protein is located in the cytoplasm, and may translocate to the nucleus upon forming the correct protein complex.
MAL8P1.153	The hypothetical protein encoded by MAL8P1.153 forms a homo-meric complex.
MAL8P1.153	This protein could be involved in protein metabolism pathways and may also cross-talk with transcription and DNA synthesis factors.
MAL8P1.153	MAL8P1.153 is expressed in sporozoite, schizont and gametocyte stages and could function in processes associated with replication and differentiation, when it might interact with different subsets of proteins.
MAL8P1.153	The product is located in the nucleus and does not bind DNA, so if it regulates transcription it must do so via interactions with other proteins.
PFL2335w	There is good evidence for this gene encoding a DNA-binding protein that regulates transcription in sporozoite and IDC stage parasites.
PFL2335w and PF10_0232 (CHD1)	The protein encoded by PFL2335w probably forms a transcription factor complex with other proteins, one of which is likely to be CHD1.

var (*Pf*EMP1)	Even though a single *var* gene is dominantly expressed there is a background level of *var* transcription, possibly including several truncated and pseudogene sequences. This may lead to a high rate of reverse transcription and reincorporation of sequences back into the genome, explaining the extraordinarily high number of pseudogenes in this family, and providing a mechanism for ensuring a reservoir of sequences from which to generate new antigenic proteins.
Species-specific, multi-copy families.	There is a relationship between the number of pseudogenes per family and the overall transcription level of the family. Furthermore, pseudogenes seem to belong almost exclusively to species-specific families, such as *var*, *rifin* and *stevor* in *P. falciparum*. Having large numbers of pseudogenes could potentially be useful for resurrecting 'dead' sequences as new proteins, thereby quickly driving evolution.
Several	Using automatic programs to assign gene function can be problematic, as demonstrated with GO functions assigned to a protein when there is a single occurrence of a particular motif or domain within a multi-domain protein. In many cases the annotation will be wrong, and this may be exacerbated in *Plasmodium* because there are many novel proteins that are likely to contain unique combinations of domains with unusual functions.

Table 6.5. Summary of novel hypotheses about gene function that emerged from exploration of *P. falciparum* functional genomic data using MaGnET as described in Chapter 6.

6.5.1 MaGnET was used to demonstrate how visualisation of functional genomic data can lead to the prediction of protein complexes

In Section 6.1 the MaGnET Expression Data Viewer was used to explore the expression patterns of a set of known and predicted CDK and cyclin genes. The results showed that several distinct patterns of co-expression involving small groups of CDKs and cyclins are clearly visible within the expression data available through MaGnET. The co-expression of a known CDK-cyclin complex (the RNA polymerase II CTD phosphorylation complex), including the CDK *Pf*mrk, the cyclin

*Pf*cyc-1 and the co-activator protein *Pf*MAT1, was clearly demonstrated in trophozoites.

MaGnET was then used to explore the expression profiles of other CDK and cyclin genes in an attempt to discover likely paired combinations based on co-expression of components. The CDK *Pf*PK5 was shown to co-express with the cyclin *Pf*cyc-4 during erythrocytic stages, so it seems likely that they will form a complex at this time (complex activity has been demonstrated *in vitro*). Also, the timing of their maximal expression is during the onset of schizogony, so they may be involved in regulating the onset of this process. Interestingly, in gametocytes, the expression profile of *Pf*PK5 was more similar to that of *Pf*cyc-2, which has not been shown to activate *Pf*PK5 in *in vitro* studies. The similarity of their expression profiles could be simply a coincidence, or it could be that they do form a functional complex in gametocytes and a co-activator protein is required. Demonstrated *in vitro* activity of *Pf*PK5 with *Pf*cyc-3 was not backed up by evidence for co-expression of their genes during the IDC or gametocyte stages. Therefore, it is unlikely that they form a functional complex during these phases of development, but they may regulate processes at other stages, such as in the mosquito, which was not investigated here. *Pf*PK5 may well form complexes with various cyclins, whose function differs between life cycle stages, and is therefore regulated by the expression of the cyclin genes.

MaGnET was also able to demonstrate that two groups, each of three CDKs and three cyclins, had distinctive co-expression profiles during the IDC. One group was maximally expressed in ring and trophozoite stage parasites and the other in schizonts. The former of these groups is likely to have a role in the regulation of the

rounds of rapid RNA and protein synthesis required for parasite growth and establishment during the early IDC. The latter group is more likely to regulate cell cycle progression during schizogony.

Sharp peaks of CDK/cyclin expression were observed during the IDC, leading to the hypothesis that rapid switch on and off of the required complex components at each phase is a mechanism for regulating the number of times a cell divides [which fits with the current model for *Plasmodium* cell cycle control: that replication continues as long as the pool of CDK/cyclins in the cytoplasm remains above a threshold level (Leete and Rubin 1996)].

The predictions of CDK/cyclin complexes made here narrows down the possible range of theoretical complexes to a sub-set of likely combinations that can be tested in the laboratory. It also provides an alternative to using computational molecular docking experiments to predict CDK/cyclin complexes, which has the advantages of being simple to use for non-bioinformaticians and not requiring knowledge of the three-dimensional structure of the proteins. In fact, the next step after the MaGnET analysis could be to run molecular docking experiments on this set of predicted CDK/cyclin complexes, utilising the comparatively modelled structures that are available within MaGnET for some of the set, in order to predict whether they may physically interact before going into the laboratory.

6.5.2 Exploration of functional genomic data using MaGnET led to new hypotheses about gene function

The principle of 'guilt by association' has been widely used to assign putative function to novel proteins based on shared similarities with groups of genes of known function in functional genomic datasets. Here, MaGnET was able to

demonstrate how this principle can be applied through use of visualisation tools to pick up on similarities across a group of genes that confer a probable functional association onto novel genes.

By examining a set of high quality (reproducible) protein-protein interactions within the Protein-Protein Interaction Viewer, it was shown how MaGnET could be used to predict the possible functional role and cellular location of novel proteins. The genes PFI1715w and MAL8P1.153 share no similarity with proteins of known function in other organisms. Both have a large number of reproducible interactions in the dataset, so were chosen as candidates for investigating whether trends in their sets of interaction partners could confer functional associations to the novel proteins.

In the case of PFI1715w, its interaction partners consisted of a large number of proteins involved in fundamental cellular processes, such as transcription, protein metabolism and cell cycle. Therefore, PFI1715w most likely encodes an intracellular protein that may be involved in novel ways of regulating cellular pathways, which are unique to the parasite. The protein encoded by MAL8P1.153 also appears to be linked to fundamental cellular processes, such as transcription, protein metabolism and DNA synthesis, and there is evidence that it is a nuclear protein and forms a homo-meric complex. Understanding how novel proteins like PFI1715w and MAL8P1.153 fit into the framework of conserved 'house-keeping' proteins involved in core processes is key to understanding how the parasite has adapted to its unique life style.

The principle of guilt by association was also used to predict DNA-binding function in a protein that has features in common with another DNA-binding protein. The protein encoded by PFL2335w shares multiple interaction partners with the

DNA-binding protein CHD1 (PF10_0232) and, moreover, they have similar expression profiles during the human stages of development, with both being particularly highly expressed in sporozoites. The hypothesis that PFL2335w also has DNA-binding function was backed up by other evidence, including the occurrence of positively charged regions in the protein sequence. This evidence leads to the further hypothesis that PFL2335w could be a binding partner for CHD1 in sporozoites, as it is known to bind to several different partners with differing functions (Hall and Georgel 2007).

6.5.3 MaGnET was successfully used to explore the properties of *P. falciparum*-specific gene families

Individual *Plasmodium* species have evolved their own specific gene families, many of which have been implicated in parasite-host interactions and immune evasion (Janssen et al. 2004). In *P. falciparum* these include *var*, *rifin* and *stevor*. Investigations of *var* gene expression profiles using the MaGnET Expression Data Viewer led to discovery of constitutively highly expressed *var*-like sequences, revealing a background level of expression that appears quite distinct from the regulated expression of a single dominant *var* gene in ring stage parasites. The *var* family is unusual in that it has many more predicted pseudogenes than all other families in the genome. It seems plausible that the high background expression level of *var* sequences may be the reason behind the large number of pseudogenes, due to processing of transcripts back into the genome. The postulated effect of this is to provide a reservoir of sequences that could potentially be brought back as 'live' sequences, thereby increasing the parasite's chances of successfully evading the host's immune response.

Further analysis of the MaGnET data revealed that there is trend for high numbers of pseudogenes occurring in species-specific gene families and that these families tend to have more abundant overall expression levels compared to other families. It would be interesting to investigate these trends further by examining the pseudogene sequences and locations within the genome to establish whether they arose by transcript processing or gene duplication. The top-three families for number of pseudogenes are *var*, *rifin* and *stevor*, the three *P. falciparum*-specific families, and the fourth is an uncharacterised family of hypothetical proteins (which also has second highest overall expression level after the *var* family). This family would seem like an important target for further investigation, since it may also have an important role in mediating host interactions, similar to the other three families.

The latter part of the investigation (correlation between number of pseudogenes per family and overall transcript abundance of the family) did not involve the MaGnET visualisation program; however, it utilised a novel combination of data from the MaGnET database. Therefore, this mini-study demonstrates how a local installation of the MaGnET software could be used by a bioinformatician. In order to recreate the same result using other resources, the researcher would need to visit PlasmoDB (Bahl et al. 2003) or GeneDB (Hertz-Fowler et al. 2004) to retrieve the family member and genomic information, and download the expression datasets from the source publications. They would then need to either manually extract the required information (very time-consuming for a whole-genome scale study), or go through significant data parsing steps in order to extract the necessary information from downloadable files. Here, just one simple script was created to extract and compare the necessary information from the database.

6.5.4 MaGnET usage simplifies the process of weeding out false annotation

Section 6.4 showed how the policy of clearly flagging up the source and reliability of predicted annotation data within MaGnET makes it easy for users to assess the information and make informed decisions about whether to trust the annotation or seek further evidence. Unfortunately, it is not possible to weed out all unreliable annotation automatically, so by providing users with the necessary factors required to make a judgement, MaGnET provides a helpful service. Other tools often do not provide a clear indication of the reliability or source of their predicted annotation, which can be misleading to biologists. An example discussed in Section 6.4.1 showed that recent efforts to increase coverage of GO terms for *P. falciparum* by automatically assigning functional terms based on occurrence of a single InterPro predicted sequence feature can be erroneous and misleading. Algorithms that take into account the global arrangement of protein domains and motifs are much more accurate and urgent effort is needed to develop this sort of annotation procedure for *P. falciparum* genes.

Furthermore, Figure 6.15 demonstrates that MaGnET displays the InterPro annotation in a different way to other tools, such as PlasmoDB and GeneDB, by vertically separating predictions on the page. This allows the user to easily compare sequence locations and attributes of hits from various sources and to establish a "consensus" opinion. Attaining a consensus from various algorithms is very important when deciding how much emphasis to place on individual domain annotations.

One promising annotation tool that bases its functional and domain assignment on global homology and consensus of multiple prediction methods is PhyloFacts (see Section 1.4.3.3) (Krishnamurthy et al. 2006). If the PhyloFacts algorithm can be applied on a large scale to *Plasmodium* genes, it might lead to more confident predicted functional annotation and, hopefully, wider coverage.

7. CONCLUSION

Overview

This thesis presented a new development in the form of a novel software tool to aid malaria biologists to explore functional genomic data about the parasite and draw hypotheses about gene function. The Introduction Chapter described recent progress in *Plasmodium* genome sequencing and functional genomics research and assessed currently available online tools for browsing and analysing the results. The field was found to be lacking resources that encouraged users to explore the functional genomic data beyond the single gene level. By comparison to similar tools available for other organisms, it was clear that a new tool providing visualisation of integrated functional genomic datasets would be useful for malaria research. The following thesis chapters described work to develop and demonstrate the use of such a tool – the Malaria Genome Exploration Tool (MaGnET).

Chapter 2 set out specific aims for software design, including user requirements and short-comings of other tools that it aimed to address. The system design included a database for local data storage and a program [Graphical User Interface (GUI)] for displaying data. Chapter 3 described the selection of publicly-available datasets, data processing and structure of the database. Chapter 4 presented the main features of the visualisation program and provided details about its implementation. This chapter also compared MaGnET to related tools, discussing specific advantages and limitations, and highlighting its novel features. Directions for potential future expansion were briefly discussed. The subsequent two chapters were dedicated to describing how MaGnET could be applied to explore functional

genomic information for individual and groups of genes, ultimately leading to new, testable hypotheses about gene function. Chapter 5 demonstrated that MaGnET could show the results of previously published studies into gene function that used other experimental and bioinformatic techniques. Chapter 6 presented the results of new analyses performed using MaGnET (and in some cases supported by other methods), which culminated in a list of new hypotheses about gene function that can be tested in the laboratory or by further bioinformatic studies.

In this chapter conclusions will be drawn about the work presented in this thesis, its strengths and weaknesses, and overall significance in its field.

7.1 Advantages of using MaGnET

MaGnET improves upon currently available graphical display facilities for various *P. falciparum* functional genomic datasets, particularly those describing protein-protein interactions and mRNA and protein expression. The main resource used by malaria biologists to access information about *Plasmodium* genes is PlasmoDB (Bahl et al. 2003). However, the visualisation provided by this resource is currently very limited, and does not extend beyond the single gene level. One of the significant advantages of MaGnET is the ability to select groups of genes at any point while browsing the data, which can be changed at any time. Other tools that have recently emerged for analysis of *Plasmodium* functional genomic datasets, such as MalPort (MalPort; http://malport.bi.up.ac.za:7070/), allow users to browse properties of pre-selected clusters of genes from microarray experiments, but not to easily alter their selection. Importantly, MaGnET allows users to carry their

selections forward between data viewers, and to save the gene list in a file for future analysis.

MaGnET consists of four integrated interfaces to the data: a Genome Viewer for display of genomic location; a Protein-Protein Interaction Viewer for display of protein interaction networks; an Expression Data Viewer for drawing mRNA and protein expression profile graphs; and a Data Analysis Viewer for querying the database and comparing functional annotation for gene lists. MaGnET also provides helpful gene fact sheets that summarise functional annotation, structural data, orthologs and paralogs, and provide link-outs to gene pages in other tools.

As well as ensuring all the sections are linked and users can carry selections between them, MaGnET also provides advanced features for integrating data-types. These include the ability to overlay different data-types, such as expression data onto genomic location and protein interaction networks. The advantage to the user is the ability to easily compare localised trends between data-types; for example, noticing strain-specific changes in gene expression over chromosomal regions.

Furthermore, MaGnET has been shown to be useful for generating novel hypotheses about function of thus-far uncharacterised genes and 'hypothetical protein encoding'genes. Most of the hypotheses that came out of analyses using MaGnET are readily testable in the laboratory or using other computational techniques. Unfortunately, due to time constraints, it was not possible to arrange for any of the hypotheses to be tested, but it would be a useful exercise for promotion of MaGnET to biologists, and could help to flag-up areas of improvement for MaGnET.

The approach to data inclusion for MaGnET also differs from other resources, because rather than include all types of data relevant to *P. falciparum,*

they are limited to the minimum necessary for understanding gene function, in order to keep MaGnET light-weight and prevent feelings of data overload by users. The datasets included were carefully selected to ensure that they would be of reasonable quality, and filtering was applied in order to try to remove the most unreliable annotation. For example, a novel pipeline was created to filter out the many low quality and redundant comparatively-modelled protein structures downloaded from ModBase (Pieper et al. 2006). Therefore, this should provide a useful service to malaria biologists, because the models in MaGnET now represent a selected set of high quality, non-redundant models.

In conclusion, MaGnET provides a novel interface for exploring *P. falciparum* functional genomic data and is useful for forming hypotheses about gene function. By ensuring that the software is freely available over the World Wide Web, and providing different access options, including browser applet and downloadable Java Web Start versions, MaGnET can be easily accessed by malaria researchers all over the world.

7.2 Limitations of the software

Due to the limited time-frame for the project, there are some types of functional genomic data that MaGnET did not address, such as metabolic pathways. Since there are plenty of tools around to visualise pathways, including the Malaria Metabolic Pathways Database (Ginsburg 2006), it was felt that leaving out this data would not detract from the overall usefulness of MaGnET. Of course, the more different types of data available, the better for forming robust hypotheses. Therefore,

finding a way to display or link to pathway data with MaGnET would be a useful expansion.

One other limitation that was highlighted by the work in Chapters 5 and 6 was the lack of statistical data within MaGnET, such as correlation scores for genes in microarray datasets. When exploring co-expressed sets of genes, it is important to have some statistical measure of how similar their expression profiles are and an indication of its significance compared to overall trends in the data. Without this data, one has to be careful when comparing gene expression profiles, because there are underlying genome-wide changes in gene expression between life cycle stages. For example, expression of a large proportion of the genome increases during the trophozoite stage when the parasite is growing and establishing itself in the erythrocyte (Gritzmacher and Reese 1984).

The lack of statistical correlation data means that MaGnET does not provide a method to search for genes that have correlated expression profiles across a time-series experiment. There are many microarray data analysis packages for performing this type of analysis, but their major limitation in respect of *P. falciparum* data is that they do not take into account local trends across a few life cycle stages. Several families of *P. falciparum* genes are known to be present at multiple life cycle stages with distinct functions (McRobert et al. 2004; Petter et al. 2007), so the ability to investigate co-expressed patterns involving sub-groups of genes over a short-time frame is essential. Recent progress to develop novel visualisation software addressing this problem has been successful for the identification of previously unseen short-term trends both generally and for subsets of genes in other organisms (Craig et al. 2005). If similar functionality can be incorporated in MaGnET, it would

provide a useful new way of exploring co-expressed genes at particular stages of the parasite's development.

MaGnET is not intended to be used as a stand-alone resource. Analyses using MaGnET should be conducted alongside other bioinformatic analyses, and it is particularly important to double-check the predicted annotation from InterPro (Mulder et al. 2007) and GO (Ashburner et al. 2000) with the source databases. Inaccuracies may have crept in, and the database versions used by InterPro may not necessarily be the most recent. MaGnET has included links to other resources, such as literature and pharmacological databases, in order to guide users to further sources of information.

In conclusion, there are some areas that need further development in MaGnET in order to ensure that it establishes a broad user-base and fulfils its potential as an important resource for malaria research. Nonetheless, MaGnET's current limitations do not detract from its contribution to the field by filling a niche for a tool facilitating exploration of *Plasmodium* functional genomic datasets.

7.3 Future outlook

MaGnET mainly includes data about *P. falciparum* genes, with some information about orthologs in other *Plasmodium* species. Several species, as well as other *P. falciparum* strains, have now had their genomes sequenced, so it would be very useful to expand MaGnET to allow for comparison between genes and non-coding regions over multiple species/strains. Comparisons do not have to be just at the genome level; comparison at the protein level can indicate conserved regions

important for function and mutations that may affect the protein, especially if they can also be mapped onto protein structure.

As already mentioned, and discussed in detail in Section 4.5.2, there are many possible directions in which MaGnET development could be taken in future. The work presented in this thesis is really just the starting point for what could become an important, comprehensive resource for malaria biologists wishing to explore *Plasmodium* functional genomic data. The difficulty is making the user community aware of the tool and how it can help their research, and encouraging them to try it out. To ensure that MaGnET continues to evolve in complement with other tools, it is necessary to gain the endorsement of major online *Plasmodium* data resources. Collaboration with the UCSC Malaria Genome Browser (Chakrabarti et al. 2007) has already been initiated, and a mechanism for directly linking into MaGnET from the genome browser is under development. Once this functionality is complete, it should be simple to set up links to MaGnET from other sources.

REFERENCES

ApiCyc - Apicomplexan Metabolic Pathways. Retrieved April 8th 2008 from http://apicyc.apidb.org/.

The Broad Institute Microbial Sequencing Centre - Plasmodium Falciparum Sequencing Project. Retrieved July 12th 2007 from http://www.broad.mit.edu/seq/msc/.

M.F. Wiser. *Cellular and Molecular Biology of Plasmodium*. Retrieved August 22nd 2008 from http://www.tulane.edu/~wiser/malaria/cmb.html.

Centers for Disease Control and Prevention - Malaria. Retrieved November 17th 2007 from http://www.cdc.gov/malaria/.

Database of Comparative Protein Structure Models. Retrieved September 7th 2007 from http://modbase.compbio.ucsf.edu/.

The Gene Ontology. Retrieved April 8th 2008 from http://www.geneontology.org/.

J. Craig Venter Institute Parasite Projects. Retrieved July 12th 2007 from http://www.tigr.org/parasiteProjects.shtml#.

JFreeChart. Retrieved April 10th 2007 from http://www.jfree.org/jfreechart/.

Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ Accessed on September 11th 2007.

B.S. Kakkilaya. *Malaria Site*. Retrieved September 18th 2008 from http://www.malariasite.com.

MalPort. Retrieved August 20th 2008 from http://malport.bi.up.ac.za:7070/.

Structural Genomics of Pathogenic Protozoa. Retrieved August 31st 2007 from http://www.sgpp.org/.

The TDR Targets Database: genomic-scale prioritization of drug targets. Retrieved August 19th 2008 from http://tdrtargets.org.

Wellcome Trust Sanger Institute Plasmodium falciparum Genome Projects. Retrieved September 3rd 2007 from http://www.sanger.ac.uk/Projects/P_falciparum/.

Wellcome Trust Sanger Institute Protozoan Genomes. Retrieved July 12th 2007 from http://www.sanger.ac.uk/Projects/Protozoa/.

World Health Organisation Malaria Fact Sheet. Retrieved July 9th 2007 from http://www.who.int/mediacentre/factsheets/fs094/en/index.html.

Al-Khedery, B., Barnwell, J.W., and Galinski, M.R. 1999. Stage-specific expression of 14-3-3 in asexual blood-stage Plasmodium. *Mol Biochem Parasitol* **102:** 117-130.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Arnot, D.E., and Gull, K. 1998. The Plasmodium cell-cycle: facts and questions. *Ann Trop Med Parasitol* **92:** 361-365.

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., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M.J., Gajria, B., Grant, G.R., Ginsburg, H., Gupta, D., Kissinger, J.C., Labo, P., et al. 2003. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res* **31:** 212-215.

Baruch, D.I., Pasloske, B.L., Singh, H.B., Bi, X., Ma, X.C., Feldman, M., Taraschi, T.F., and Howard, R.J. 1995. Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82:** 77-87.

Bastien, O., Aude, J.C., Roy, S., and Marechal, E. 2004. Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. *Bioinformatics* **20:** 534-537.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340:** 783-795.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Bernstein, H.J. 2000. Recent changes to RasMol, recombining the variants. *Trends Biochem Sci* **25:** 453-455.

Berry, A.E., Gardner, M.J., Caspers, G.J., Roos, D.S., and Berriman, M. 2004. Curation of the Plasmodium falciparum genome. *Trends Parasitol* **20:** 548-552.

Birkholtz, L.M., Bastien, O., Wells, G., Grando, D., Joubert, F., Kasam, V., Zimmermann, M., Ortet, P., Jacq, N., Saidani, N., et al. 2006. Integration and mining of malaria molecular, functional

and pharmacological data: how far are we from a chemogenomic knowledge space? *Malar J* **5:** 110.

Bonnefoy, S., Guillotte, M., Langsley, G., and Mercereau-Puijalon, O. 1992. Plasmodium falciparum: characterization of gene R45 encoding a trophozoite antigen containing a central block of six amino acid repeats. *Exp Parasitol* **74:** 441-451.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. 2003. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol* **1:** E5.

Bracchi-Ricard, V., Barik, S., Delvecchio, C., Doerig, C., Chakrabarti, R., and Chakrabarti, D. 2000. PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from Plasmodium falciparum. *Biochem J* **347 Pt 1:** 255-263.

Breitkreutz, B.J., Stark, C., and Tyers, M. 2003. Osprey: a network visualization system. *Genome Biol* **4:** R22.

Carlton, J., Silva, J., and Hall, N. 2005. The genome of model malaria parasites, and comparative genomics. *Curr Issues Mol Biol* **7:** 23-37.

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., et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* **419:** 512-519.

Carret, C.K., Horrocks, P., Konfortov, B., Winzeler, E., Qureshi, M., Newbold, C., and Ivens, A. 2005. Microarray-based comparative genomic analyses of the human malaria parasite Plasmodium falciparum using Affymetrix arrays. *Mol Biochem Parasitol* **144:** 177-186.

Carter, R., and Mendis, K.N. 2002. Evolutionary and historical aspects of the burden of malaria. *Clin Microbiol Rev* **15:** 564-594.

Castillo-Davis, C.I., Bedford, T.B., and Hartl, D.L. 2004. Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. *Mol Biol Evol* **21:** 1422-1427.

Chakrabarti, K., Pearson, M., Grate, L., Sterne-Weiler, T., Deans, J., Donohue, J.P., and Ares, M., Jr. 2007. Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *Rna* **13:** 1923-1939.

Chen, F., Mackey, A.J., Stoeckert, C.J., Jr., and Roos, D.S. 2006a. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34:** D363-368.

Chen, Y., Jirage, D., Caridha, D., Kathcart, A.K., Cortes, E.A., Dennull, R.A., Geyer, J.A., Prigge, S.T., and Waters, N.C. 2006b. Identification of an effector protein and gain-of-function mutants that activate Pfmrk, a malarial cyclin-dependent protein kinase. *Mol Biochem Parasitol* **149:** 48-57.

Cheng, Q., Cloonan, N., Fischer, K., Thompson, J., Waine, G., Lanzer, M., and Saul, A. 1998. stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. *Mol Biochem Parasitol* **97:** 161-176.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26:** 183-186.

Conway, D.J., Fanello, C., Lloyd, J.M., Al-Joubori, B.M., Baloch, A.H., Somanath, S.D., Roper, C., Oduola, A.M., Mulder, B., Povoa, M.M., et al. 2000. Origin of Plasmodium falciparum malaria is traced by mitochondrial DNA. *Mol Biochem Parasitol* **111:** 163-171.

Coppel, R.L., Roos, D.S., and Bozdech, Z. 2004. The genomics of malaria infection. *Trends Parasitol* **20:** 553-557.

Coulson, R.M., Hall, N., and Ouzounis, C.A. 2004. Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. *Genome Res* **14:** 1548-1554.

Cowman, A.F., and Crabb, B.S. 2006. Invasion of red blood cells by malaria parasites. *Cell* **124:** 755-766.

Craig, P., Kennedy, J., and Cumming, A. 2005. Animated interval scatter-plot views for the exploratory analysis of large-scale microarray time-course data. *Information Visualization* **4:** 149-163.

Date, S.V., and Stoeckert, C.J., Jr. 2006. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. *Genome Res* **16:** 542-549.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* **34:** W362-365.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4:** P3.

Doerig, C. 2005. Protein kinases regulating Plasmodium proliferation and development. In *Molecular Approaches to Malaria*. (ed. I.W. Sherman), pp. 542. America Society for Microbiology.

Doerig, C., Doerig, C., Horrocks, P., Coyle, J., Carlton, J., Sultan, A., Arnot, D., and Carter, R. 1995. Pfcrk-1, a developmentally regulated cdc2-related protein kinase of Plasmodium falciparum. *Mol Biochem Parasitol* **70:** 167-174.

Doerig, C., Endicott, J., and Chakrabarti, D. 2002. Cyclin-dependent kinase homologues of Plasmodium falciparum. *Int J Parasitol* **32:** 1575-1585.

Doolittle, R.F. 2002. The grand assault. *Nature* **419:** 493-494.

Ettwiller, L., and Paten, B. 2004. Functional genomics: Guilt by multiple association. *Heredity* **92:** 481-482.

Fan, Q., An, L., and Cui, L. 2004. Plasmodium falciparum histone acetyltransferase, a yeast GCN5 homologue involved in chromatin remodeling. *Eukaryot Cell* **3:** 264-276.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res* **34:** D247-251.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. 2008. Ensembl 2008. *Nucleic Acids Res* **36:** D707-714.

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., et al. 2002. A proteomic view of the Plasmodium falciparum life cycle. *Nature* **419:** 520-526.

Galinski, M.R., Dluzewski, A.R., and Barnwell, J.W. 2005. A mechanistic approach to merozoite invasion of red blood cells: merozoite biogenesis, rupture, and invasion of erythrocytes. In *Molecular Approaches to Malaria*. (ed. I.W. Sherman), pp. 113-168. American Society for Microbiology.

Gardiner, D.L., Dixon, M.W., Spielmann, T., Skinner-Adams, T.S., Hawthorne, P.L., Ortega, M.R., Kemp, D.J., and Trenholme, K.R. 2005. Implication of a Plasmodium falciparum gene in the switch between asexual reproduction and gametocytogenesis. *Mol Biochem Parasitol* **140:** 153-160.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419:** 498-511.

Ginsburg, H. 2006. Progress in in silico functional genomics: the malaria Metabolic Pathways database. *Trends Parasitol* **22:** 238-240.

Goddard, N.H., Cannon, R.C., and Howell, F.W. 2003. Axiope tools for data management and data sharing. *Neuroinformatics* **1:** 271-284.

Gowthaman, R., Sekhar, D., Kalita, M.K., and Gupta, D. 2005. A database for Plasmodium falciparum protein models. *Bioinformation* **1:** 50-51.

Greenwood, B.M., Bojang, K., Whitty, C.J., and Targett, G.A. 2005. Malaria. *Lancet* **365:** 1487-1498.

Gritzmacher, C.A., and Reese, R.T. 1984. Protein and nucleic acid synthesis during synchronized growth of Plasmodium falciparum. *J Bacteriol* **160:** 1165-1167.

Gunasekera, A.M., Myrick, A., Militello, K.T., Sims, J.S., Dong, C.K., Gierahn, T., Le Roch, K., Winzeler, E., and Wirth, D.F. 2007. Regulatory motifs uncovered among gene expression clusters in Plasmodium falciparum. *Mol Biochem Parasitol* **153:** 19-30.

Gysin, J., Gavoille, S., Mattei, D., Scherf, A., Bonnefoy, S., Mercereau-Puijalon, O., Feldmann, T., Kun, J., Muller-Hill, B., and Pereira da Silva, L. 1993. In vitro phagocytosis inhibition assay for the screening of potential candidate antigens for sub-unit vaccines against the asexual blood stage of Plasmodium falciparum. *J Immunol Methods* **159:** 209-219.

Hall, J.A., and Georgel, P.T. 2007. CHD proteins: a diverse family with strong ties. *Biochem Cell Biol* **85:** 463-476.

Hall, N., and Carlton, J. 2005. Comparative genomics of malaria parasites. *Curr Opin Genet Dev* **15:** 609-613.

Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., et al. 2005. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307:** 82-86.

Harrison, P.M., and Gerstein, M. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol* **318:** 1155-1174.
Hawthorne, P.L., Trenholme, K.R., Skinner-Adams, T.S., Spielmann, T., Fischer, K., Dixon, M.W., Ortega, M.R., Anderson, K.L., Kemp, D.J., and Gardiner, D.L. 2004. A novel Plasmodium falciparum ring stage protein, REX, is located in Maurer's clefts. *Mol Biochem Parasitol* **136:** 181-189.
Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., et al. 2004. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res* **32:** D339-343.
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., et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298:** 129-149.
Horrocks, P., Kyes, S.A., Bull, P.C., and Deitsch, K.W. 2005. Molecular aspects of antigenic variation in *Plasmodium falciparum*. In *Molecular Approaches to Malaria*. (ed. I.W. Sherman), pp. 399-415. American Society for Microbiology.
Hyland, C., Pinney, J.W., McConkey, G.A., and Westhead, D.R. 2006. metaSHARK: a WWW platform for interactive exploration of metabolic networks. *Nucleic Acids Res* **34:** W725-728.
Janssen, C.S., Phillips, R.S., Turner, C.M., and Barrett, M.P. 2004. Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Res* **32:** 5712-5720.
Johnson, J.E., Stromvik, M.V., Silverstein, K.A., Crow, J.A., Shoop, E., and Retzel, E.F. 2003. TableView: portable genomic data visualization. *Bioinformatics* **19:** 1292-1293.
Joubert, Y., and Joubert, F. 2008. A structural annotation resource for the selection of putative target proteins in the malaria parasite. *Malar J* **7:** 90.
Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34:** D354-357.
Karp, P.D., Paley, S., and Romero, P. 2002. The Pathway Tools software. *Bioinformatics* **18 Suppl 1:** S225-232.
Kaviratne, M., Khan, S.M., Jarra, W., and Preiser, P.R. 2002. Small variant STEVOR antigen is uniquely located within Maurer's clefts in Plasmodium falciparum-infected red blood cells. *Eukaryot Cell* **1:** 926-935.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996-1006.
Kidgell, C., Volkman, S.K., Daily, J., Borevitz, J.O., Plouffe, D., Zhou, Y., Johnson, J.R., Le Roch, K., Sarr, O., Ndir, O., et al. 2006. A systematic map of genetic variation in Plasmodium falciparum. *PLoS Pathog* **2:** e57.
Kim, H.J., Song, E.J., Lee, Y.S., Kim, E., and Lee, K.J. 2005. Human Fas-associated factor 1 interacts with heat shock protein 70 and negatively regulates chaperone activity. *J Biol Chem* **280:** 8125-8133.
Koegl, M., and Uetz, P. 2007. Improving yeast two-hybrid screening systems. *Brief Funct Genomic Proteomic* **6:** 302-312.
Kooij, T.W., Carlton, J.M., Bidwell, S.L., Hall, N., Ramesar, J., Janse, C.J., and Waters, A.P. 2005. A Plasmodium whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathog* **1:** e44.
Krishnamurthy, N., Brown, D.P., Kirshner, D., and Sjolander, K. 2006. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biol* **7:** R83.
LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., et al. 2005. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* **438:** 103-107.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.
Lanzer, M., de Bruin, D., Wertheimer, S.P., and Ravetch, J.V. 1994. Transcriptional and nucleosomal characterization of a subtelomeric gene cluster flanking a site of chromosomal rearrangements in Plasmodium falciparum. *Nucleic Acids Res* **22:** 4176-4182.

Lanzer, M., Wickert, H., Krohne, G., Vincensini, L., and Braun Breton, C. 2006. Maurer's clefts: a novel multi-functional organelle in the cytoplasm of Plasmodium falciparum-infected erythrocytes. *Int J Parasitol* **36:** 23-36.

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., et al. 2002. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* **419:** 537-542.

Law, P.J., Claudel-Renard, C., Joubert, F., Louw, A.I., and Berger, D.K. 2008. MADIBA: a web server toolkit for biological interpretation of Plasmodium and plant gene clusters. *BMC Genomics* **9:** 105.

Le Roch, K., Sestier, C., Dorin, D., Waters, N., Kappes, B., Chakrabarti, D., Meijer, L., and Doerig, C. 2000. Activation of a Plasmodium falciparum cdc2-related kinase by heterologous p25 and cyclin H. Functional characterization of a P. falciparum cyclin homologue. *J Biol Chem* **275:** 8952-8958.

Le Roch, K.G., Johnson, J.R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S.F., Williamson, K.C., Holder, A.A., Carucci, D.J., et al. 2004. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res* **14:** 2308-2318.

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., et al. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301:** 1503-1508.

Leete, T.H., and Rubin, H. 1996. Malaria and the cell cycle. *Parasitol Today* **12:** 442-444.

Li, J.L., Robson, K.J., Chen, J.L., Targett, G.A., and Baker, D.A. 1996. Pfmrk, a MO15-related protein kinase from Plasmodium falciparum. Gene cloning, sequence, stage-specific expression and chromosome localization. *Eur J Biochem* **241:** 805-813.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178-2189.

Li, X., Kim, J., Zhou, J., Gu, W., and Quigg, R. 2005. Use of signal thresholds to determine significant changes in microarray data analyses. *Genet. Mol. Biol.* **28:** 191-200.

Llinas, M., Bozdech, Z., Wong, E.D., Adai, A.T., and DeRisi, J.L. 2006. Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res* **34:** 1166-1173.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McLaren, P., North, P., et al. 2007. FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol* **8:** R129.

Marmorstein, R., and Berger, S.L. 2001. Structure and function of bromodomains in chromatin-regulating complexes. *Gene* **272:** 1-9.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29:** 291-325.

Marti, M., Good, R.T., Rug, M., Knuepfer, E., and Cowman, A.F. 2004. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306:** 1930-1933.

Martz, E. 2002. Protein Explorer: easy yet powerful macromolecular visualization. *Trends Biochem Sci* **27:** 107-109.

McDermott, J., Guerquin, M., Frazier, Z., Chang, A.N., and Samudrala, R. 2005. BIOVERSE: enhancements to the framework for structural, functional and contextual modeling of proteins and proteomes. *Nucleic Acids Res* **33:** W324-325.

McGowan, C.H. 2003. Regulation of the eukaryotic cell cycle. *Prog Cell Cycle Res* **5:** 1-4.

McRobert, L., Preiser, P., Sharp, S., Jarra, W., Kaviratne, M., Taylor, M.C., Renia, L., and Sutherland, C.J. 2004. Distinct trafficking and localization of STEVOR proteins in three stages of the Plasmodium falciparum life cycle. *Infect Immun* **72:** 6597-6602.

Mehlin, C. 2005. Structure-based drug discovery for Plasmodium falciparum. *Comb Chem High Throughput Screen* **8:** 5-14.

Mehlin, C., Boni, E., Buckner, F.S., Engel, L., Feist, T., Gelb, M.H., Haji, L., Kim, D., Liu, C., Mueller, N., et al. 2006. Heterologous expression of proteins from Plasmodium falciparum: results from 1000 genes. *Mol Biochem Parasitol* **148:** 144-160.

Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci* **11:** 430-448.

Merckx, A., Le Roch, K., Nivez, M.P., Dorin, D., Alano, P., Gutierrez, G.J., Nebreda, A.R., Goldring, D., Whittle, C., Patterson, S., et al. 2003. Identification and initial characterization of three

novel cyclin-related proteins of the human malaria parasite Plasmodium falciparum. *J Biol Chem* **278:** 39839-39850.
Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F., and Trajanoski, Z. 2005. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res* **33:** W633-637.
Mrowka, R. 2001. A Java applet for visualizing protein-protein interaction. *Bioinformatics* **17:** 669-671.
Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., et al. 2007. New developments in the InterPro database. *Nucleic Acids Res* **35:** D224-228.
Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247:** 536-540.
Noble, M.E., Endicott, J.A., Brown, N.R., and Johnson, L.N. 1997. The cyclin box fold: protein recognition in cell-cycle and transcription control. *Trends Biochem Sci* **22:** 482-487.
Nunes, M.C., Goldring, J.P., Doerig, C., and Scherf, A. 2007. A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell. *Mol Microbiol* **63:** 391-403.
Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH--a hierarchic classification of protein domain structures. *Structure* **5:** 1093-1108.
Orton, R.J. 2006. Visualising and exploring linked functional genomic data sets in the Yeast Exploration Tool Integrator (YETI). PhD thesis. School of Biology, University of Edinburgh, Edinburgh. Accessed on 08/11/2007.
Orton, R.J., Sellers, W.I., and Gerloff, D.L. 2004. YETI: Yeast Exploration Tool Integrator. *Bioinformatics* **20:** 284-285.
Perraut, R., Marrama, L., Diouf, B., Fontenille, D., Tall, A., Sokhna, C., Trape, J.F., Garraud, O., and Mercereau-Puijalon, O. 2003. Distinct surrogate markers for protection against Plasmodium falciparum infection and clinical malaria identified in a Senegalese community after radical drug cure. *J Infect Dis* **188:** 1940-1950.
Peters, J., Fowler, E., Gatton, M., Chen, N., Saul, A., and Cheng, Q. 2002. High diversity and rapid changeover of expressed var genes during the acute phase of Plasmodium falciparum infections in human volunteers. *Proc Natl Acad Sci U S A* **99:** 10689-10694.
Petter, M., Haeggstrom, M., Khattab, A., Fernandez, V., Klinkert, M.Q., and Wahlgren, M. 2007. Variant proteins of the Plasmodium falciparum RIFIN family show distinct subcellular localization and developmental expression patterns. *Mol Biochem Parasitol* **156:** 51-61.
Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., et al. 2006. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **34:** D291-295.
Ralph, S.A. 2005. The apicoplast. In *Molecular Approaches of Malaria*. (ed. I.W. Sherman), pp. 272-289. American Society for Microbiology.
Rees, C.A., Demeter, J., Matese, J.C., Botstein, D., and Sherlock, G. 2004. GeneXplorer: an interactive web application for microarray data visualization and analysis. *BMC Bioinformatics* **5:** 141.
Rogerson, S.J. 2003. Sequestration: causes and consequences. *Redox Rep* **8:** 295-299.
Ross-Macdonald, P.B., Graeser, R., Kappes, B., Franklin, R., and Williamson, D.H. 1994. Isolation and expression of a gene specifying a cdc2-like protein kinase from the human malaria parasite Plasmodium falciparum. *Eur J Biochem* **220:** 693-701.
Rowe, J.A. 2005. Rosetting. In *Molecular Approaches to Malaria*. (ed. I.W. Sherman), pp. 416-426. American Society for Microbiology.
Ryu, S.W., Lee, S.J., Park, M.Y., Jun, J.I., Jung, Y.K., and Kim, E. 2003. Fas-associated factor 1, FAF1, is a member of Fas death-inducing signaling complex. *J Biol Chem* **278:** 24003-24010.
Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A., and Quackenbush, J. 2006. TM4 microarray software suite. *Methods Enzymol* **411:** 134-193.
Salomonis, N., Hanspers, K., Zambon, A.C., Vranizan, K., Lawlor, S.C., Dahlquist, K.D., Doniger, S.W., Stuart, J., Conklin, B.R., and Pico, A.R. 2007. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* **8:** 217.

Schneider, A.G., and Mercereau-Puijalon, O. 2005. A new Apicomplexa-specific protein kinase family: multiple members in Plasmodium falciparum, all with an export signature. *BMC Genomics* **6:** 30.

Shah, N.H., and Fedoroff, N.V. 2004. CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics* **20:** 1196-1197.

Shock, J.L., Fischer, K.F., and DeRisi, J.L. 2007. Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol* **8:** R134.

Sinden, R.E., and Gilles, H.M. 2002. The malaria parasites. In *Essential Malariology*, 4th ed. (eds. D.A. Warrell, and H.M. Gilles), pp. 8-34. Arnold.

Smith, J.D., Chitnis, C.E., Craig, A.G., Roberts, D.J., Hudson-Taylor, D.E., Peterson, D.S., Pinches, R., Newbold, C.I., and Miller, L.H. 1995. Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82:** 101-110.

Song, E.J., Yim, S.H., Kim, E., Kim, N.S., and Lee, K.J. 2005. Human Fas-associated factor 1, interacting with ubiquitinated proteins and valosin-containing protein, is involved in the ubiquitin-proteasome pathway. *Mol Cell Biol* **25:** 2511-2524.

Spielmann, T., Hawthorne, P.L., Dixon, M.W., Hannemann, M., Klotz, K., Kemp, D.J., Klonis, N., Tilley, L., Trenholme, K.R., and Gardiner, D.L. 2006. A cluster of ring stage-specific genes linked to a locus implicated in cytoadherence in Plasmodium falciparum codes for PEXEL-negative and PEXEL-positive proteins exported into the host cell. *Mol Biol Cell* **17:** 3613-3624.

Stubbs, J., Simpson, K.M., Triglia, T., Plouffe, D., Tonkin, C.J., Duraisingh, M.T., Maier, A.G., Winzeler, E.A., and Cowman, A.F. 2005. Molecular mechanism for switching of P. falciparum invasion pathways into human erythrocytes. *Science* **309:** 1384-1387.

Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A., and Wellems, T.E. 1995. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82:** 89-100.

Thompson, J., Janse, C.J., and Waters, A.P. 2001. Comparative genomics in Plasmodium: a tool for the identification of genes and functional analysis. *Mol Biochem Parasitol* **118:** 147-154.

Triglia, T., Thompson, J.K., and Cowman, A.F. 2001. An EBA175 homologue which is transcribed but not translated in erythrocytic stages of Plasmodium falciparum. *Mol Biochem Parasitol* **116:** 55-63.

Vaidya, A.B. 2005. The mitochondrion. In *Molecular Approaches to Malaria*. (ed. I.W. Sherman), pp. 234-252. American Society for Microbiology.

van Dooren, G.G., Stimmler, L.M., and McFadden, G.I. 2006. Metabolic maps and functions of the Plasmodium mitochondrion. *FEMS Microbiol Rev* **30:** 596-630.

van Noort, V., and Huynen, M.A. 2006. Combinatorial gene regulation in Plasmodium falciparum. *Trends Genet* **22:** 73-78.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vincensini, L., Richert, S., Blisnick, T., Van Dorsselaer, A., Leize-Wagner, E., Rabilloud, T., and Braun Breton, C. 2005. Proteomic analysis identifies novel proteins of the Maurer's clefts, a secretory compartment delivering Plasmodium falciparum proteins to the surface of its host cell. *Mol Cell Proteomics* **4:** 582-593.

Waller, K.L., McBride, S.M., Kim, K., and McDonald, T.V. 2008. Characterization of two putative potassium channels in Plasmodium falciparum. *Malar J* **7:** 19.

Waller, R.F., and McFadden, G.I. 2005. The apicoplast: a review of the derived plastid of apicomplexan parasites. *Curr Issues Mol Biol* **7:** 57-79.

Ward, P., Equinet, L., Packer, J., and Doerig, C. 2004. Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC Genomics* **5:** 79.

Watanabe, J., Suzuki, Y., Sasaki, M., and Sugano, S. 2004. Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, Plasmodium species. *Nucleic Acids Res* **32:** D334-338.

Wickham, M.E., Rug, M., Ralph, S.A., Klonis, N., McFadden, G.I., Tilley, L., and Cowman, A.F. 2001. Trafficking and assembly of the cytoadherence complex in Plasmodium falciparum-infected human erythrocytes. *Embo J* **20:** 5636-5649.

Wilson, R.J., Denny, P.W., Preiser, P.R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D.J., Moore, P.W., et al. 1996. Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum. *J Mol Biol* **261:** 155-172.

Winter, G., Kawai, S., Haeggstrom, M., Kaneko, O., von Euler, A., Kawazu, S., Palm, D., Fernandez, V., and Wahlgren, M. 2005. SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. *J Exp Med* **201:** 1853-1863.

Wirth, D.F. 2002. Biological revelations. *Nature* **419:** 495-496.

Yeh, I., Hanekamp, T., Tsoka, S., Karp, P.D., and Altman, R.B. 2004. Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res* **14:** 917-924.

Young, J.A., Fivelman, Q.L., Blair, P.L., de la Vega, P., Le Roch, K.G., Zhou, Y., Carucci, D.J., Baker, D.A., and Winzeler, E.A. 2005. The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* **143:** 67-79.

Zehetner, G. 2003. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res* **31:** 3799-3803.

Zhou, Y., Young, J.A., Santrosyan, A., Chen, K., Yan, S.F., and Winzeler, E.A. 2005. In silico gene function prediction using ontology-based pattern identification. *Bioinformatics* **21:** 1237-1245.

APPENDICES

Appendix A

A list of tables in the MaGnET database with a short description of each column.

'CHROMOSOMES' table

Field	Description	Type
id*	A unique identifier assigned to the chromosome by the MaGnET program	Integer
species	The *Plasmodium* species to which the chromosome belongs	Text
strain	The *Plasmodium species* strain to which the chromosome belongs	Text
chr	The name/number of the chromosome	Text
length	The length of the chromosome in base pairs	Integer
sequence	The nucleotide sequence of the chromosome (currently empty due to lack of storage space)	Text

'GENES' table

Field	Description	Type
magnet_id*	A unique identifier given to the gene by the MaGnET program	Integer
gene_id	The standard gene identifier assigned by the sequencing centre	Text
alias	Any previous identifiers this gene was known by	Text
type	Type of gene e.g. protein coding	Text
strand	The chromosome strand on which the gene is found (c or w)	Character
chr	The chromosome where the gene is found	Text
species	The *Plasmodium* species to which the gene belongs	Text
strain	The *Plasmodium species* strain to which the gene belongs	Text
keywords	Keywords describing the name/function of the gene	Text
product_name	The name given to the product of the gene	Text
start	The starting coordinate of the gene on the strand	Integer
end	The ending coordinate of the gene on the strand	Integer
length	The total length in base pairs of the gene	Integer
protein_sequence	The amino acid sequence of the gene's protein product	Text
nucleotide_sequence	The gene's nucleotide sequence	Text
num_of_exons	The number of exons the gene has	Integer
curation1, curation2, curation3, curation4	Annotation provided by the sequencing consortium	Text

'EXONS' table

Field	Description	Type
ex_id*	A unique identifier given to the exon by the MaGnET program	Integer
magnet_id	The magnet_id of the gene the exon belongs to	Integer
exon_number	The exon number within the gene	Integer
exon_start	The starting coordinate of the exon on the strand	Integer
exon_end	The ending coordinate of the exon on the strand	Integer
exon_length	The length of the exon in base pairs	Integer

'ORTHOLOGUES' table

Field	Description	Type
cluster_id*	A unique identifier given to the cluster	Text
pknowlesi	The number of cluster members in the *P. knowlesi* genome	Integer
pberghei	The number of cluster members in the *P. berghei* genome	Integer
pchabaudi	The number of cluster members in the *P. chabaudi* genome	Integer
pvivax	The number of cluster members in the *P. vivax* genome	Integer
pf3d7	The number of cluster members in the *P. falciparum 3D7* genome	Integer
pyoelii	The number of cluster members in the *P. yoelii* genome (currently not available)	Integer
cluster_members	The standard gene identifiers of the cluster members	Text

'DOMAINS' table

Field	Description	Type
feature_id*	A unique identifier given to the domain/ sequence feature	Integer
gene_id	The standard gene identifier assigned by the sequencing centre	Text
type	The type of domain/ sequence feature or the method used to predict it	Text
domain_id	A unique identifier for the category of domain/ sequence feature (assigned by the database that made the prediction)	Text
description	A description of the domain/ sequence feature	Text
start	The start position in the protein sequence	Integer
end	The end position in the protein sequence	Integer
evalue	The expectation (E) value of the prediction	Text
date	The date the prediction was made	Text
interpro	A unique identifier for the category of domain/ sequence feature assigned by InterPro	Text
note	Additional information about the domain/ sequence feature	Text

'GENE_FEATURES' table

Field	Description	Type
feature_id*	A unique identifier given to the feature	Integer
gene_id	The standard gene identifier assigned by the sequencing centre	Text
signal_peptide	The probability of this being a signal peptide sequence	Text
signal_anchor	The probability of this being a signal anchor sequence	Text
cleavage_site	The probability of there being a cleavage site	Text
coordinates	The location in the protein sequence	Text
note	Additional information about the prediction	Text
type	The prediction method	Text

'GO_DATA' table

Field	Description	Type
ont_id*	A unique identifier given to the Gene Ontogology annotation by the MaGnET program	Integer
magnet_id	The magnet_id of the gene the annotation is assigned to	Integer
gene_id	The standard gene identifier assigned by the sequencing centre	TextA
aspect	The Gene Ontology aspect (C = cellular component, P = biological process, F = molecular function)	Text
go_id	The Gene Ontology term identifier	Text
term_name	The Gene Ontology term name	Text
evidence_tag	The evidence tag associated with this annotation	Text
reference	The reference for this annotation	Text
evidence_with	Contains the database and identifier of contributing sequence if annotation was inferred by sequence similarity, or contributing GO identifier if annotation was inferred from another GO annotation.	Text
date	The date on which the annotation was made	Text

'GENE_ONTOLOGIES' table

Field	Description	Type
term_id*	The Gene Ontology term identifier	Text
term_name	The Gene Ontology term name	Text
aspect	The Gene Ontology aspect (C = cellular component, P = biological process, F = molecular function)	Text
alt_id	An alternative (obsolete) Gene Ontology identifier for this term	Text
description	A more detailed description of the term	Text

'PDB_STRUCTURES' table

Field	Description	Type
pdb_struct_id*	A unique identifier assigned to the structure by the MaGnET program	Integer
pdb_code	The PDB unique four character identifier	Text
chain	The chain identifier within the protein that this structure corresponds to	Text
magnet_id	The magnet_id of the gene encoding the protein this structure represents	Integer
gene_id	The standard gene identifier of the gene encoding the protein this structure represents	Text
sequence	The amino acid sequence of the solved structure	Text
start	The starting position of the solved structure	Integer
end	The ending position of the solved structure	

'STRUCTURE_MODELS' table

Field	Description	Type
model_id*	A unique identifier assigned to the model structure by the ModBase database	Text
magnet_id	The magnet_id of the gene encoding the protein this structure represents	Integer
model_seq	The amino acid sequence of the model	Text
gene_id	The standard gene identifier of the gene encoding the protein this structure represents	Text

seq_id	The sequence identity of the modelled sequence to its template sequence	Double
model_score	A numerical score assigned to the model by the ModBase database	Double
e_value	The expectation (E) value of the match between the modelled sequence and its template sequence	Double
template_pdb	The four character PDB identifier of the template structure	Text
template_chain	The chain identifier of the template structure	Text
target_length	The length of the modelled sequence	Integer
target_begin	The starting residue number of the model	Integer
target_end	The ending residue number of the model	Integer
template_begin	The starting residue of the template structure	Integer
template_end	The ending residue of the template structure	Integer
date	The date the model was deposited in the ModBase database	Text
run	A name given to the specific program run in which the model was created	Text
note	Any additional information about the model	Text

'INTERACTIONS' table

Field	Description	Type
int_id*	A unique identifier given to the interaction by the MaGnET program	Integer
bait_orf	The standard gene identifier of the gene encoding the bait protein	Text
bait_magnet_id	The magnet_id of the gene encoding the bait protein	Integer
prey_orf	The standard gene identifier of the gene encoding the prey protein	Text
prey_magnet_id	The magnet_id of the gene encoding the prey protein	Integer
independent_searches	The number of independent searches in which this interaction was observed	Integer
times_observed	The total number of times this interaction was observed	Integer
prey_no_of_bait	The number of prey proteins that interact with this bait protein	Integer
bait_no_of_prey	The number of bait proteins that interact with this prey protein	Integer
bait_unique_ints	The total number of unique interactions in which this bait protein participates	Integer
prey_unique_ints	The total number of unique interactions in which this prey protein participates	Integer
interaction_type	The type of interaction that is occuring e.g. self, reciprocal	Text
study	The name of the study in which the interaction was observed	Text

'PROTEIN_EXP_STUDY_LASONDER_2002' table

Field	Description	Type
gene_id*	The standard gene identifier of the gene	Text
Troph_and_Schiz, Gametocytes, Gametes	The gene's expression level at these timepoints	Double

'PROTEIN_EXP_STUDY_LEROCH_2004' table

Field	Description	Type
gene_id*	The standard gene identifier of the gene	Text
Gamete, Gameto, Mero, Ring, Troph, Schiz, Sporo	The gene's expression level at these timepoints	Double

'MRNA_EXP_STUDY_LEROCH_2003' table

Field	Description	Type
gene_id*	The standard gene identifier of the gene	Text
Early_ring_S, Late_ring_S, Early_trophozoite_S, Late_trophozoite_S, Early_schizont_S, Late_schizont_S, Merozoite_S, Early_ring_T, Late_ring_T, Early_trophozoite_T, Late_trophozoite_T, Early_schizont_T, Late_schizont_T, Merozoite_T, Gametocyte, Sporozoite	The gene's expression level at these timepoints	Double

'MRNA_EXP_STUDY_YOUNG_2005' table

Field	Description	Type
gene_id*	The standard gene identifier of the gene	Text
Sporozoite, Early_ring_S, Late_ring_S, Early_trophozoite_S, Late_trophozoite_S, Early_schizont_S, Late_schizont_S, Merozoite_S, Early_ring_T, Late_ring_T, Early_trophozoite_T, Late_trophozoite_T, Early_schizont_T, Late_schizont_T, Merozoite_T, Gametocyte_3D7_Early_Day_1, Gametocyte_3D7_Early_Day_2, Gametocyte_3D7_Early_Day_3, Gametocyte_3D7_Early_Day_4, Gametocyte_3D7_Day_1, Gametocyte_3D7_Day_2, Gametocyte_3D7_Day_3, Gametocyte_3D7_Day_6, Gametocyte_3D7_Day_8, Gametocyte_3D7_Day_12, Gametocyte_NF54_Day_1, Gametocyte_NF54_Day_2, Gametocyte_NF54_Day_3, Gametocyte_NF54_Day_4, Gametocyte_NF54_Day_5, GametocyteNF54_Day_6, Gametocyte_NF54_Day_7, Gametocyte_NF54_Day_8, Gametocyte_NF54_Day_9, Gametocyte_NF54_Day_10, Gametocyte_NF54_Day_11, Gametocyte_NF54_Day_12, Gametocyte_NF54_Day_13	The gene's expression level at these timepoints	Double

'MRNA_EXP_STUDY_LLINAS_HB3_QC_2006' table

Field	Description	Type
oligo*	A unique identifier assigned to the oligonucleotide	Text
gene_id	The standard gene identifier of the gene	Text

'MRNA_EXP_STUDY_LLINAS_DD2_QC_2006' table

Field	Description	Type
oligo*	A unique identifier assigned to the oligonucleotide	Text
gene_id	The standard gene identifier of the gene	Text

'MRNA_EXP_STUDY_LLINAS_3D7_QC_2006' table

Field	Description	Type
oligo*	A unique identifier assigned to the oligonucleotide	Text
gene_id	The standard gene identifier of the gene	Text

'MRNA_DECAY_SHOCK_2007' table

Field	Description	Type
oligo*	A unique identifier assigned to the oligonucleotide	Text
gene_id	The standard gene identifier of the gene	Text
ring_0min	The gene's expression ratio at the start of the measurements in the ring life cycle stage	Double
ring_240min	The gene's expression ratio at the end of the measurements (after 240 minutes) in the ring life cycle stage	Double
ring_half_life	The half life of the mRNA during the ring life cycle stage	Double
troph_0min	The gene's expression ratio at the start of the measurements in the trophozoite life cycle stage	Double
troph_240min	The gene's expression ratio at the end of the measurements (after 240 minutes) in the trophozoite life cycle stage	Double
troph_half_life	The half life of the gene during the trophozoite life cycle stage	Double
schiz_0min	The gene's expression ratio at the start of the measurements in the schizont life cycle stage	Double
schiz_240min	The gene's expression ratio at the end of the measurements (after 240 minutes) in the schizont life cycle stage	Double
schiz_half_life	The half life of the mRNA during the schizont life cycle stage	Double
late_schiz_0min	The gene's expression ratio at the start of the measurements in the late schizont life cycle stage	Double
late_schiz_240min	The gene's expression ratio at the end of the measurements (after 240 minutes) in the late schizont life cycle stage	Double
late_schiz_half_life	The half life of the mRNA during the late schizont life cycle stage	Double

* This is the table's primary key

Appendix B

Details of the datasets that contribute data to the MaGnET database, listing the species and strains of *Plasmodium* to which they relate, the online sources they were obtained from – including their URLs, their dates of release and references to relevant publications.

Dataset	Organism	Source	URL	Version/ release date	Reference
Chromosome sequences (nuclear chr 1-14)	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/*	2.1.4 09/07/2007	(Gardner et al. 2002)
Gene sequences (nuclear chr 1-14)	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/*	2.1.4 10/07/2007	(Gardner et al. 2002)
Protein sequences (nuclear chr 1-14)	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/*	2.1.4 10/07/2007	(Gardner et al. 2002)
Sequencing centre genome annotation (nuclear chr 1-14)	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/*	2.1.4 09/07/2007	(Gardner et al. 2002)
RNA genes	*P. falciparum* 3D7	Ares laboratory at UCSC	*http://areslab.ucsc.edu/*	September 2007	(Chakrabarti et al. 2007)
Apicoplast chromosome sequence	*P. falciparum* C10	PlasmoDB	*http://www.plasmodb.org/common/downloads/release-5.4/PfalciparumPlastid/*	5.4 24/09/2007	(Wilson et al. 1996)
Gene sequences					
Protein sequences					
Additional annotation		NCBI nucleotide database	*http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore*	23/05/2005	

Mitochondrial chromosome sequence	*P. falciparum* NF54	PlasmoDB	*http://www.plasmodb.org/common/downloads/release-5.4/PfalciparumMito/*	5.4 24/09/2007	(Conway et al. 2000)
Gene sequences					
Protein sequences					
Additional annotation		NCBI nucleotide database	*http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore*	24/11/2000	
Gene Ontology annotation	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/gene_association_file*	03/04/2008	(Ashburner et al. 2000)
Gene Ontology term descriptions	N/A	The Gene Ontology	*http://www.geneontology.org/GO.downloads.ontology.shtml*	08/04/2008	
Ortholog and paralog groupings	*P. falciparum* 3D7, *P. vivax*, *P. berghei*, *P.chabaudi*, *P.knowlesi*	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/*	24/10/2006	(Li et al. 2003)
InterPro sequence features	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/interpro*	07/12/2007	(Mulder et al. 2007)
SignalP predicted signal peptides	*P. falciparum* 3D7	Wellcome Trust Sanger Institute	*ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7*	2.1.4 09/07/2007	(Bendtsen et al. 2004)
Protein-protein interaction network from yeast two-hybrid study	*P. falciparum* 3D7	Journal publication	Supplementary material	3/11/2005	(LaCount et al. 2005)
Experimentally-solved 3D protein structures	*P. falciparum* 3D7	RCSB Protein Data Bank	*http://www.rcsb.org/pdb*	20/11/2007	(Berman et al. 2000)
Comparatively-modelled 3D protein structures	*P. falciparum* 3D7	ModBase	*http://modbase.compbio.ucsf.edu/*	Nov 2007	(Pieper et al. 2006)

MRNA expression data for intra-erythrocytic developmental cycle	*P. falciparum* 3D7, Dd2 and HB3	DeRisi laboratory at UCSF	*http://malaria.ucsf.edu/comparison/*	February 2006	(Llinas et al. 2006)
MRNA expression data for sexual development stages	*P. falciparum* 3D7 and NF54	Winzeler laboratory at Scripps Research Institute	*http://carrier.gnf.org/publications/Gametocyte/*	September 2005	(Young et al. 2005)
MRNA expression data for several life cycle stages	*P. falciparum* 3D7	Winzeler laboratory at Scripps Research Institute	*http://carrier.gnf.org/publications/CellCycle/*	September 2003	(Le Roch et al. 2003)
MRNA decay profiles	*P. falciparum* 3D7	Journal publication	Supplementary material	August 2007	(Shock et al. 2007)
Protein expression data for several life cycle stages	*P. falciparum* 3D7	Journal publication	Supplementary material	November 2004	(Florens et al. 2002; Le Roch et al. 2004)
Protein expression data for a few life cycle stages	*P. falciparum* NF54	Journal publication	Supplementary material	October 2002	(Lasonder et al. 2002)

Appendix C

List of database update programs written to facilitate automatic data processing and MaGnET database population. They can be used whenever new versions of source files are released (provided the file format has not changed). Most are standalone Java programs, with some Perl scripts.

File	Type	Purpose
AreslabGeneUpdater	Java program	To update RNA gene predictions downloaded from the UCSC Malaria Genome Browser.
CreateSingleFASTAFile	Java program	To create a single FASTA sequence file from multiple individual sequence files. Required for PDB structure processing.
DatabaseConnector	Java class	Used to establish a connection to the MaGnET database. Required by all other Java programs.
ExpDataGeneFilter	Java program	To filter out any oligonucleotides from a microarray dataset that do not currently map to gene models. (For use with spotted cDNA-type arrays)
FilterOutLowLevelGenesFromExpData	Java program	Not currently used. Filters out all genes whose expression never rises above a particular cut-off level. (For use with Affymetrix-type arrays)
MapOldIDsToNew	Java program	Update gene ids that have changed in a new release of the genome
UpdateExpressionStudyTables	Java program	Add a new mRNA or protein expression study dataset
UpdateMRNADecayData	Java program	Add a new mRNA decay study dataset
UpdateTableChromosomes	Java program	Update chromosome data
UpdateTableDomains	Java program	Update InterPro annotation
UpdateTableGO_DATA	Java program	Update GO annotation
UpdateTableGeneOntologies	Java program	Updates the GO term names and descriptions associated with GO numbers

UpdateTableGenesAndExons	Java program	Updates gene data, including genomic location, exon boundaries, product names and sequencing centre annotation
UpdateTableInteractions	Java program	Add a new Y2H protein-protein interaction dataset
UpdateTableOrthologues	Java program	Update OrthoMCL-generated gene families
UpdateTablePDBStructures	Java program	Update details of experimental 3-D protein structures
The following programs are all used for filtering out low quality, redundant comparative protein structure models:		
CreateDatabaseIDFile	Java class	Creates a file of URLs pointing to ModBase model sets for individual genes
CreateGeneIDList	Java class	Creates a list of gene ids required by the above program
FetchModels	Java program	Calls the above two classes
formatclusters_winv.pl	Perl script	Formats the BLASTCLUST output into a more useful format
Model	Java class	A class to represent and hold information about a particular model
ModelMatch	Java class	A class to represent and hold information about a match between two models generated by the BLASTCLUST program
pdbres2seqmodall.pl	Perl script	Calculates amino acid sequences from PDB format structures files and prints them out in FASTA format
PopulateTableStructuralModelsPartOne	Java program	Populates a database table with model information
PopulateTableStructuralModelsPartTwo	Java program	Creates new structure coordinate files for the set of filtered models
RemoveRedundantModelsPartOne	Java program	Reads output from round one of BLASTCLUST analysis and removes redundant models from the database table
RemoveRedundantModelsPartTwo	Java program	Reads output from round two of BLASTCLUST analysis and removes further redundant models from the database table
runblastclust_myversion_part1.pl	Perl script	Runs the BLASTCLUST program against a database of model amino acid sequences for round one of redundant model filtering

runblastclust_myversion_part2.pl	Perl script	Runs the BLASTCLUST program against a database of model amino acid sequences for round two of redundant model filtering
SequenceTable	Java class	Used by some of the above programs to hold a list of gene ids and corresponding model ids
SortModels	Java program	Performs initial filtering of the models to remove models not meeting quality control criteria, such as minimum sequence identity to template structure