Transcriptional regulation of yir genes in Plasmodium yoelii yoelii infected erythrocytes

Thesis

UNRESTRICTED

Transcriptional regulation of yir genes in Plasmodium yoelii yoelii infected erythrocytes

Jannik Fonager, MSc

Division of Parasitology
National Institute for Medical Research
London, United Kingdom

March 2006

A thesis submitted in part fulfilment of the requirements of the Open University for the degree of Doctor of Philosophy.

[illegible]

Abstract

In 1998, a large multigene family was discovered in *Plasmodium vivax* (del Portillo et al., 1998). This multigene family was termed *vir*, and later studies showed that *vir* homologues existed in the three rodent malaria species, *Plasmodium chabaudi (cir), Plasmodium berghei (bir)* and *Plasmodium yoelii (yir).* By 5x coverage sequencing of *Plasmodium yoelii* 17X, 838 *yir* genes were predicted (Carlton et al., 2002), making this the largest known multigene family in *Plasmodium.* YIR proteins are expressed at the surface of infected erythrocytes (Cunningham et al., 2005), and therefore this family is thought to be involved in antigenic variation.

The aim of this thesis was to examine how *P.yoelii* regulates transcription of the *yir* family. The *yir* gene structure was verified experimentally, and phylogenetic analysis showed that *yir* genes could be divided into five supergroups consisting of *yir* genes with different sizes and subtelomeric localisation. In the blood stages, numerous *yir* genes were transcribed from all the supergroups in immunocompromised mice. However, a maximum of two *yir* genes were transcribed in single infected erythrocytes at the Schizont stage, which suggested strong silencing mechanisms. The transcriptional start and polyadenylation sites were identified experimentally, and it was found that both occurred at highly conserved motifs. In addition, the transcription initiation site was located close to an unusual and universally conserved triple-repeat motif, and it was found that all *yir* transcripts in two populations of parasites initiated downstream of this motif. Transfection experiments were performed in order to examine the role of this motif, but no solid conclusions could be drawn from these. Several alternative splicing events were detected in the *yir* 5′UTRs, and one of these led to exon 1 skipping of a *yir* gene. Through bioinformatic analysis of *yir* 5′ intergenic regions, it was found that the UTR introns had a discrete distribution amongst the *yir* supergroups.

Acknowledgements

The work presented in this thesis, I carried out at the National Institute for Medical Research, London, UK between October 2002 and August 2005.

First, I would like to thank my supervisors Dr Peter Preiser and Dr Jean Langhorne for inviting me to work with the *yir* genes, and for always encouraging and supporting me throughout this project. I would also like to thank Dr William Jarra for being a fountain of knowledge in parasite biology, and for teaching me how to perform sequencing and micromanipulation of parasites, as well as helping me with the transfection experiments. Thanks to Dr Deirdre Cunningham for proofreading this thesis and to Dr Delmiro Fernandes-Reyes for constructive criticism of Chapter IV. I would also like to thank Dr Deirdre Cunningham for teaching me how to perform Northern blotting and for always taking an interest in my project and for a pleasant atmosphere in the Lab. PhD student Sandra Koernig I would like to thank for helping me with the FACS analysis and for being a capable and helpful fellow PhD student on the *yir* genes. Past and present members of the Langhorne lab are thanked for constructive debates and for a nice atmosphere on the row as well as for the occasional social arrangements.

I would like to thank my mother for caring support during this project. Finally, I am especially grateful to Kirsten for her unending care and support and for reminding me of a world outside of the lab.

Abbreviations

ABS: Asexual blood stage

aRNA: anti-sense RNA

BLAST: Basic Local Alignment Search Tool

Bp: Base pairs

cDNA: Complementary DNA (Reverse transcribed RNA)

Contigs: Contiguous set of overlapping DNA sequences

gDNA: Genomic DNA

iRBC: Infected red blood cell

I.P: Intraperitoneal

IRES: Internal Ribosomal Entry Site

LD-PCR: Long distance PCR

ME-tree: Maximum evolution tree

MEME: Multiple Em for Motif Elicitation

MP-tree: Maximum parsimony tree

NJ-tree: Neighbour joining tree

NT: Nucleotides (refers to positions on a single stranded RNA molecule)

ORF: Open reading frame

PCR: Polymerase chain reaction

RACE: Rapid amplification of cDNA ends

RBC: Red blood cell

RT: Reverse transcriptase

UPGMA-tree: Unweighted Pair Group Method with Arithmetic Mean-tree

UTR: Untranslated region

Contents

Chapter I Introduction

Chapter II Materials and Methods

Chapter III Gene structure

Chapter IV YIR phylogeny

Chapter V Transcription profile

Chapter VI Single cell RT-PCR

Chapter VII Characterisation of yir UTRs

Chapter VIII Transfection experiments

Chapter IX Alternative splicing

Chapter X Conclusion and future perspectives

Supplementary material

On the supplied CD-ROM, the supplementary figures referred to in the text, are located. These figures are:

S2.1: Sequence of the *Pbtubα*-II vector

S2.2: Sequence of IR1019

S4.1: Alignment of consensus sequences

S8.1: Alignment between IR1019 and the three best matching contigs

S8.2: Sequencing of inserts

S8.3: Transfection experiment 1 (parasitaemia curves)

S8.4: Transfection experiment 2 (parasitaemia curves)

S9.1: Sv 1 splice type

S9.2: Sv 2 splice type

S9.3: Sv 3 splice type

S9.4: Non-spliced transcript

In addition, the CD-ROM contains branched (DNA NJ, 5` intergenic region UPGMA and 3` intergenic region NJ) trees, to allow for a closer inspection of bootstrap values. An excel spreadsheet with the *yir* groups and supergroups is also included.

A CD-rom containing the supplementary information
is embedded in the back inside cover of this thesis.

Chapter I

Introduction

1.1 Malaria

Malaria is the world's most prevalent infectious disease, responsible for 300-500 million annual cases resulting in 1-2.7 million deaths (Phillips, R.S et al., 2001 and WHO). Up to 90% of all annual clinical episodes occur in Africa alone and it is estimated that 10% of all disability-adjusted life-years in Africa are due to malaria (WHO and Breman et al., 2004). In Africa, 75% of all deaths caused by malaria occur in children under the age of 5 (Breman et al., 2001) and in total, malaria is responsible for 20% of all deaths among children in this age-group (WHO). Typified by Uganda and according to the Ugandan Ministry of Health (http://www.health.go.ug/malara.htm), 18 to 37 out of every 1000 children under the age of 5 dies of malaria in low and high endemic regions respectively (between 70000 and 110000 children). In addition, poor families spend up to 25% of their income on malaria treatment and prevention and on average 7 working days are lost per malaria episode. As 58% of malaria episodes occur in the poorest 20% of the world's population (Breman et al., 2004), it has a significant impact on the economic development of affected countries. It has been estimated to cost 12 billion US$ in lost GDP in primarily sub-Saharan Africa (Phillips, R.S et al., 2001 and WHO) and retards economic growth by 1.3% (Gallup and Sachs et al., 2001). As if these facts were not enough, malaria is on the increase due to the spread of chloroquine and sulfadoxine-pyrimethamine resistant parasite strains (WHO).

Four species of *Plasmodium* infect man (*Plasmodium falciparum, vivax, ovale* and *malariae*). It has been proposed in a recent study that the common ancestor to *P. falciparum* and *P. vivax* existed some 200000 to 300000 years ago, and that this coincides with the emergence of modern man (Jongwutiwes et al., 2005). *P.vivax* can be phylogenetically linked to *Plasmodium* species, infecting macaque monkeys in Asia, although the existence of Duffy negative blood groups (which protects against *P. vivax* infections) in Africa makes this suggested regional origin less clear (Escalante et al., 2005). Modern *P. falciparum* is thought to have evolved in Africa some 6000 years ago, following an expansion of malaria populations (infecting humans) some 10,000 years ago (Joy et al., 2003). Several human traits, such as sickle cell formation and the Duffy negative blood group confer resistance to malaria, and it is thought that the co-existence of *Plasmodium* and *Homo sapiens* has had a strong influence on human evolution (as reviewed by Kwiatkowski et al., 2005).

Symptoms of malaria are caused by parasites in the asexual erythrocytic stage and occur between 7 to 30 days (dependent on *Plasmodium* species, according to http://www.cdc.gov/malaria/disease.htm#incubation) after infection through a mosquito bite. The severity of the infection depends on *Plasmodium* species and past infection history of the patient. Infection with *P.falciparum* can have the most severe outcome, if untreated. Malaria infections can lead to the following clinical features in adults (according to http://www.who.int/malaria/docs/hbsm_adults.htm): cerebral malaria (CM), anaemia, renal failure, hypoglycaemia, fluid and electrolyte disturbances, pulmonary oedema, circulatory collapse (algid malaria), abnormal bleeding and disseminated intravascular coagulation, high fever and malarial haemoglobinuria. Many of these clinical features also occur in children, which is the most susceptible age group, as 75% of all deaths from malaria occurs in children under the age of 5 (Breman et al., 2001). Especially CM has the worst prognosis, as even with treatment the mortality is 30% (English et. al., 2002). A total of 7% of children surviving CM are left with permanent neurological problems (http://www.rbm.who.int/cmc_upload/0/000/015/367/RBMInfosheet_6.htm).

Malaria infections during pregnancy (placental malaria) leads to low birth weight and premature delivery, both of which can cause neonatal death and impaired cognitive development.

1.2 Malaria phylogeny

The malaria parasite, *Plasmodium*, belongs to the Protista kingdom and the phylum Apicomplexa. Seven parasitic genera are present within the Apicomplexan phylum (*Plasmodium, Toxoplasma, Cryptosporidium, Babesia, Isospora, Cyclospora* and *Sarcocystis*), and a common feature among them is the existence of the apical complex, which is believed to be responsible for the attachment and penetration of the host cell membrane.

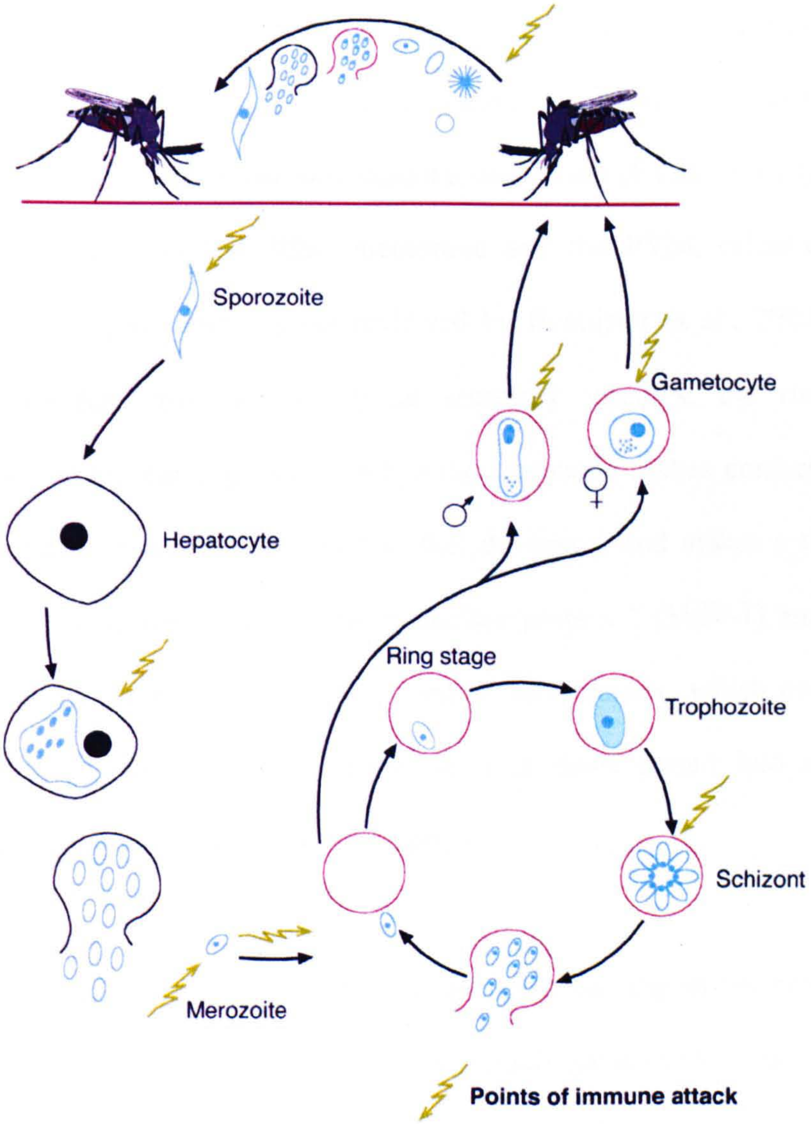
1.3 The biology of malaria

The malaria parasite has a complex life cycle involving both an invertebrate mosquito vector of the genus *Anopheles* and a vertebrate host. Figure 1.1 shows the malaria life cycle, which begins with the injection of sporozoites into the host bloodstream. The sporozoites migrate quickly to the liver where they invade hepatocytes and replicate. The resulting liver stage merozoites are then released back into the bloodstream, which initiates the asexual blood stage cycle by invasion of erythrocytes. Some malaria species and clones are able to invade normocytes, whereas others prefer to invade reticulocytes. In the asexual blood stage, the malaria parasites undergo a series of distinct developmental stages inside the infected erythrocyte. The first stage is the ring stage where the parasite begins to feed on haemoglobin. As the ring stage parasite develops further, it begins to export proteins to the infected erythrocyte (iRBC) surface, which allow the iRBC to adhere to

Figure 1.1

The malaria lifecycle

The malaria lifecycle begins with the bite of an infected female *Anopheline spp.* mosquito (shown left above the line). Sporozoites, injected by the mosquito, quickly migrate to the hepatocyte cells of the liver. Here, they undergo several rounds of replication, resulting in the release of merozoites into the host bloodstream. In the bloodstream, the merozoites invade erythrocytes and develop through the ring and trophozoite stages into the multinucleated schizont containing many merozoites. At the end of the schizont stage, the infected erythrocyte ruptures releasing merozoites into the bloodstream. This initiates a new cycle of erythrocyte invasion and replication of parasites. Eventually some parasites develop into the sexual stage gametocytes, which are taken up by a mosquito. In the mosquito, male and female gametocytes emerge from their host cells, forming male and female gametes respectively. In the mosquito midgut, the male gamete fertilizes the female gamete, leading to the diploid zygote. The zygote then develops into the motile ookinete, which passes through the midgut wall and develops into the oocyst. After rupture of the oocyst cell wall, sporozoites emerge and make their way to the salivary gland, where they are ready to infect a human again when the mosquito bites. At several points during the life cycle, the parasite is exposed to the host's immune system (indicated by arrows in the figure).


Sporozoite
Hepatocyte
Merozoite
Gametocyte
♂
♀
Ring stage
Trophozoite
Schizont
Points of immune attack

endothelial cells and other RBCs (as reviewed by Bannister et al., 2003). The ring eventually develops into a trophozoite, which is the stage where most feeding and growth occurs. Several proteins are exported to the surface of the iRBC in this stage, including *Plasmodium falciparum* erythrocyte membrane protein 1 (Pfemp1), which allows the iRBC to adhere to endothelial cells and withdraw from circulation. The next stage is called the schizont stage, and is characterised by a series of nuclear divisions resulting in a species dependent number of merozoites located within a membrane called the parasitophorous vacuolar membrane (PVM). Finally a protease dependent process lyses the iRBC membrane and the PVM, releasing the free merozoites into the bloodstream (as reviewed by Bannister et al., 2003). The free merozoites contain three sets of apical secretory vesicles: the rhoptries, the micronemes and the dense granules. When the merozoite makes contact with a red blood cell (RBC), it reorientates itself so that the apical end makes a tight contact with the RBC membrane. The merozoite surface protein 1 (MSP-1) and the apical membrane antigen 1 (AMA-1) are involved in this process, which ends with the internalisation of the merozoite and the onset of development into a ring stage parasite (as reviewed by Bannister et al., 2003).

Environmental factors, which are not yet understood, can trigger the onset of sexual development in the vertebrate host. Male and female gametocyte (micro and macro-gametes respectively) development is not caused by the presence of specific sex chromosomes, and the decision to produce either male or female gametocytes has been taken prior to the schizont stage (Silvestrini et al., 2000). Male and female gametocytes differ in which kinases and phosphatases they express (Khan et al., 2005). As the gametocytes remain in a dormant state in the vertebrate host after the initial stages of development, it is thought that the kinases respond to different environmental signals in the mosquito, and are needed to continue the sexual differentiation. In the mosquito, fertilisation occurs leading to the production of sporozoites from the oocysts, which through a mosquito bite are transferred to another vertebrate host, thus completing the cycle.

1.4 Encounters with the immune system during invasion

For *P. falciparum* approximately 3700 sporozoites and for *P. vivax* approximately 3400 sporozoites are produced in the oocysts in the midgut of the mosquito, and it was estimated that approximately 20% of these eventually reach the salivary glands (Rosenberg et al., 1991). Between 10 and 100 *P. falciparum* sporozoites are injected per bite (Rosenberg et al., 1990 and Ponnudurai et al., 1991), and an average of 123 for the rodent malaria *P.yoelii* (Medica et al., 2005). In this stage, called the pre-erythrocytic stage, the parasite expresses a number of surface proteins necessary for the liver stage invasion. Antibodies against several of these: circumsporozoite protein (CSP), Liver stage antigen-1 (LSA-1) and thrombospondin-related adhesive protein (TRAP) are found in natural infections, and high levels of antibodies correlated with protection (John et al., 2005). After the amplification in the liver, the parasite expresses another set of surface proteins necessary for erythrocyte invasion. The most notable of these are Apical Membrane antigen (AMA-1), Merozoite Surface Protein 1 and 2 (MSP-1 and MSP-2), and antibodies against these are also found in natural infections (John et al., 2005; Omosun et al., 2005 and Polley et al., 2005). Of these, especially MSP-1 has received special attention as a possible vaccine candidate. MSP-1 is proteolytically processed at the surface of the merozoite, and antibodies that inhibit this processing also inhibit invasion (Blackman et al., 1990). However, in human sera, naturally occurring antibodies block the binding of inhibitory antibodies, and this is speculated to be a mechanism of immune evasion (as reviewed by Holder et al., 1999).

One very important question is why does the host's antibody responses not lead to complete immunity to malaria. One reason could be that both the invasive stages are of a very short duration, and for the hepatocyte, invasion only occurs once during an infection. Therefore, if a high titre level of antibodies or other immune mechanisms are not operating at the time of sporozoite injection, the sporozoites will reach a safe haven inside the hepatocytes before the immune system has detected their presence. The erythrocyte invasion is repeated multiple times during an infection, so the immune system should have an opportunity to react to the free merozoites. However, at this stage, the number of merozoites released is so large, that even a high titre level of antibodies and other highly active immune mechanisms might not be sufficient to prevent reinvasion. In addition, the blocking antibodies against MSP-1 could be one mechanism, where the parasite guides the immune response against a functionally unimportant part of the protein, and thereby prevents the binding of antibodies that would seriously interfere with it's function.

1.5 The asexual blood stages

During the asexual blood stages, the parasite is mostly located within erythrocytes apart from the short time it takes for the merozoites to reinvade a new erythrocyte. Ideally it would be assumed that once inside the erythrocyte, the parasite is protected from the host's immune system if it can avoid signalling its existence to the immune system. However, the presence of parasite antigens on the surface of *Plasmodium*-infected erythrocytes was first discovered in the 1930's by Eaton (as reviewed by Kyes et al., 2001). At that time it was discovered that serum from a monkey infected with *P.knowlesi* was able to agglutinate infected, but not uninfected cells. This became known as the Schizont Infected Cell Agglutination test (SICA). In the years that followed, further studies showed that the SICA phenomenon was caused by polymorphically different parasite proteins on the surface of the infected erythrocyte (as reviewed by Kyes et al., 2001). In the 75 years since then, several other polymorphic surface expressed parasite proteins have been identified in most, if not all, of the studied *Plasmodium* species. The latest discovery was the finding of the SURFIN family in 2005 (Winter et al., 2005).

1.6 Multigene families encode variant antigens on the iRBC surface

Today it is known that several multigene families exist in the genome of the parasite, and since several of these encode predicted transmembrane regions, they could possibly be expressed at the iRBC surface. With the advent of near to complete genomic sequencing of several *Plasmodium* species, the last couple of years have led to a more accurate picture of the copy numbers and species distribution of various multigene families. The complete genome sequence of *P.falciparum* revealed that this parasite strain (3D7) contains 59 *var*, 149 *rif*, 28 *stevor*, 13 *etramp* and 10 *surfin* genes among others not mentioned here (Gardner et al., 2002, Spielman et al., 2003 and Winter et al., 2005). A common trait for these multigene families is that they are mostly located in subtelomeric regions of the chromosomes. The subtelomeric location of these genes is particularly important as this is a region that has been shown to have a high rate of recombination, leading to the generation of new diversity within genes (Freitas-Junior et al., 2000). However some restraints on which genes can recombine exist and this is proposed to be related to functional conservation of some groups of genes, while others are allowed to diverge further (Kraemer et al., 2003).

1.7 Why express anything on the surface at all?

There have been various suggestions to why the parasite would express foreign proteins at all on the host cell surface in the asexual blood stage, as this at first glance would be an unnecessary exposure to the host's immune system. However it is possible that these proteins serve important biological functions for the parasite while it is located inside the erythrocyte. Some of the most prominent suggestions for what the proteins could be doing are (as reviewed by Kyes et al., 2001):

1 Parasitized erythrocytes are destroyed in the spleen. By expressing proteins that would mediate adherence to host surface receptors, the parasite could avoid passing through the spleen.

2 If the parasite population grows too rapidly, it would kill the host before transmission can occur. Therefore some parasites become suicidal and flag their presence to the immune system in order to control the overall growth rate.

3 Hiding of senescence signals on the iRBC. The parasite might damage the iRBC in a way that makes it appear senescent to the mechanisms that normally remove aged erythrocytes. The function of the iRBC surface proteins could be to mask the senescence signals displayed at the iRBC surface. This explanation could be grouped together with explanation number 1.

4 Immunomodulation. By expressing variant surface molecules, the parasite might interfere with the efficiency of antigen presenting cells, which would in turn decrease the efficiency of an adaptive immune response directed against the parasite.

In the possible explanations above it is important to distinguish between what is cause and what is effect. A primary cause for the existence of surface proteins could be any of the three first explanations and in addition it is also possible that these proteins have other important biological functions related to nutrient uptake for instance. No matter what the primary cause is, the proteins on the iRBC surface are potentially targeted by the host's adaptive immune system. As an effect of this, the parasite would have to avoid the host's adaptive immune system. It is also possible that the function of some surface proteins is to divert an immune response away from a biologically crucial surface molecule that for various functional reasons cannot vary enough to constantly avoid the immune system on its own. Therefore antigenic variation in the blood stages would intuitively always be linked to the maintenance of an important biological function either directly or indirectly. Explanation two is the exception, as this would suggest that surface proteins are deliberately put on the surface to be recognized. If this was the case, a few gene copies with a high immunogenic potential could achieve this. Instead, highly diverse members of surface proteins are expressed in a sequential fashion with very little cross reactivity between the antibodies that recognize each type (Newbold et al., 1992 and Bull et al., 1999). Although this does not absolutely disproves that growth control could be achieved through this method, the generally held idea is that the purpose of antigenic variation is to avoid the immune system whilst at all time maintain a variant protein on the surface which plays an important role for the parasites interaction with its surroundings.

1.8 The proposed function of Pfempl

The var genes encode Pfempl proteins. From the sequenced *P.falciparum* strain, 59 *var* genes have been identified (Gardner et al., 2002). The function of the Pfempl proteins have been studied in great detail, and from this it has been found that they are responsible for sequestration to host surface receptors in tissue or on other erythrocytes (rosetting). This function would make Pfemp1 responsible for the withdrawal from circulation, and be consistent with the need for the parasite to avoid passing through the spleen.

The spleen is a complex organ that is able to remove damaged iRBCs and contains pathogen-specific T and B cells (as reviewed by Engwerda et al., 2004). It has been proposed that the parasite damages the iRBC in a way that resembles ageing, which the spleen reacts to by removing it from circulation (as reviewed by Sherman et al., 2004). When blood enters the spleen, its flow is reduced in a region called the marginal zone. It has been proposed that this increases the efficiency with which the macrophages and dendritic cells can capture parasite antigens (as reviewed by Engwerda et al., 2004). Some early studies showed that the presence of a functional spleen was needed to induce withdrawal from circulation and also lead to the emergence of new variant antigens on the iRBC surface in monkey and rodent malaria models (Hommel et al., 1983, Lean S.A et al., 1982; Gilks et al., 1990). The withdrawal from circulation and the emergence of new surface antigens can now be explained by the Pfemp1 (*var*) model in *P.falciparum*. Initially it was found that Pfemp1 undergoes clonal antigenic variation (Biggs et al., 1991), and that defined antigenic and adhesive Pfemp1 phenotypes were found to switch to other types at a rate of 2.4% per generation in the absence of immune pressure (Roberts et al., 1992). However in a more recent *in vivo* study, an 18% switching rate was found among *var* transcripts (Gatton et al., 2003). This discrepancy could be caused by the presence of the host's immune mechanisms in the latter study and in addition, only a limited subset of adhesive Pfempl proteins were used to estimate the switching rate in the first study.

It is now known that specific Pfempl proteins are able to bind to a variety of host surface receptors, including: CD36, ICAM-1 and VCAM, CSA, IgM, Blood group antigens A and B (Baruch et al., 1995; Smith et al., 1995 and Newbold et al., 1997 and as reviewed by Flick et al., 2004). The interactions with the various host receptors are mediated through three major domains (DBL, CIDR and C2) in the Pfempl proteins (Chen et al., 2000 and Gardner et al., 2002). Certain combinations of these domains exist for the different Pfempl proteins (Gardner et al., 2002), whereas others that are theoretically possible are not found (Kraemer et al., 2003) and this possibly reflects some functional limitations. The Pfempl proteins are therefore responsible for the withdrawal from circulation and thereby enabling the parasite to avoid passing through the spleen. The numerous copies of the *var* genes allows the parasite to adhere to a wide variety of host receptors and also enables it to switch to other variants if an immune response is directed against a particular type.

1.9 Pfempl expression

How the Pfempl proteins switches between different variants has been the focus of intense studies, and it was found that at the ring stage, between 1 to 15 different *var* transcripts are present, followed by allelic exclusion of all but one type as the parasite develops into a trophozoite (Fernandez et al., 2002 and Scherf et al., 1998 and Chen et al., 1998). The single trophozoite *var* transcript encodes for a Pfempl protein with a distinct adhesive phenotype (Scherf et al., 1998). In the early ring stage, some *var* transcripts occur more often than others, however as the ring develops, a mosaic-like pattern of *var* transcription emerges (Fernandez et al., 2002).

Only one full-length transcript is present in the ring stages, whereas the remainder are truncated (Taylor et al., 2000). *Var* transcription occurs *in situ* and *var* genes on some chromosomes seemed have a higher likelihood of being transcribed (Scherf et al., 1998 and Fernandez et al., 2002). This was indirectly supported by the finding that different *var* genes have distinct and reproducible on and off rates, suggesting that a hard-wired mechanism governs *var* transcription (Horrocks et al., 2004). From these studies, it is now known that Pfemp1 expression is regulated at the transcriptional level where several *var* variants are transcribed in the ring stages, followed by a commitment to only one of these transcripts in the trophozoite stage. How the *var* genes are regulated at the transcriptional level has also been intensively investigated and will be covered in detail later in this chapter.

1.10 The RIF and STEVOR families

Rif and *stevor* genes are located in clusters together with the *var* genes at the chromosome ends (Gardner et al., 2002). *Rif* and *stevor* are two multigene families that belong to the same superfamily (Cheng et al., 1998). The *rif* genes are slightly larger than the *stevor* genes and both families contain two predicted transmembrane domains separated by a hyper-variable loop, which is thought to face outwards from the membrane (Cheng et al., 1998, Sam-Yellowe et al., 2004 and as reviewed by Blythe et al., 2004). Naturally occurring antibodies against RIF proteins has been found in human infections (Kyes et al., 1999) and a high level of RIF antibodies is inversely correlated with the severity of an infection (Abdel-Latif et al., 2003). It was earlier thought that RIF proteins were responsible for the clustering together of erythrocytes (rosetting) in *P.falciparum* infections, and therefore they were called rosettins. However, it was found that at best the RIF proteins only played a minor role in this process compared with Pfemp (Kyes et al., 1999), which has also later been supported by the finding that the DBL domain in Pfemp is involved in the rosette formation (Chen et al., 2004). Both RIF and STEVOR proteins were found to contain targeting motifs (e.g. PEXEL) in their amino acid sequences (Marti et al., 2004 and Hiller et al., 2004). Unlike the proposed localisation of RIF on the erythrocyte surface, STEVOR proteins seem to be located in a structure beneath the erythrocyte membrane called the Maurer's cleft in the asexual blood stages (Kaviratne et al., 2002). The Maurer's cleft structure is thought to be important for the sorting of proteins to the erythrocyte surface (Przyborski et al., 2005). However STEVOR protein were found to localize independently of the Maurer's cleft structure in the gametocyte stages, suggesting a different function for STEVOR in this stage (Mc Robert et al., 2004). It has been proposed that the function of STEVOR in the asexual blood stages is to protect the Maurer's cleft in the late schizont stages where the erythrocyte becomes permeable to soluble antibodies (Blythe et al., 2004), whereas RIF could play a similar role earlier in the asexual blood stages where the erythrocyte is not permeable. This is partly supported by fact that *stevor* transcription is detected later in the asexual blood stages than *rif* transcription (Kaviratne et al., 2002). In a recent microarray analysis of *in vivo* samples from patients infected with *P. falciparum*, it was found that the only significantly up regulated group of genes with a GO function, were genes encoding proteins exported to the iRBC surface when compared to *in vitro* samples. Both *stevor* and *rif* were found to be the two most abundant transcripts (Daily et al., 2005). Up to three different *stevor* transcripts can be detected in single iRBCs, with one very abundant transcript found in several iRBCs (Kaviratne et al., 2002), but apart from this, virtually nothing is known about how the *rif* and *stevor* genes are regulated.

1.11 The Py235 multigene family

In the rodent malaria, *P.yoelii*, extensive studies have been performed on the subtelomerically located *Py235* multigene family with 14 genomic copies (Owen et al., 1999 and Carlton et al., 2002). The Py235 protein has homologous in several malaria species and is located at the rhoptry of merozoites and is involved in the invasion of erythrocytes (as reviewed by Gruner et al., 2004 and Khan et al., 2001). As both a non-lethal and a lethal *P.yoelii* strain (17XNL and YM respectively) exist, it has been proposed that Py235 could be involved in the different outcome of an infection through the shift from a predominantly reticulocyte preference in the non-lethal strain to a normocyte preference in the lethal strain (Khan et al., 2001). Although the *Py235* repertoire is the same in both the non-lethal and the lethal strain, the lethal strain was found to transcribe a more limited subset of *Py235* genes (Preiser et al., 1998). Different *Py235* transcripts were found in different stages of the parasite's life cycle and antibodies to Py235 reduced the infectivity of liver stage sporozoites, which strongly suggest that this family is involved in both hepatocyte and erythrocyte invasion through regulated expression of different types in the different life stages (Preiser et al., 2002). Strong transcriptional control exists for this multigene family, as it was found that upon the trophozoite to schizont transition, individual merozoites within the schizont each transcribes only one *Py235* gene in a clonally different manner (Preiser et al., 1999).

1.12 *Plasmodium* chromosome ends

The multigene families described in the preceding sections have three things in common: location of the protein products on the iRBC surface; few genes transcribed per iRBC, and subtelomeric location of the majority of the genes. Apart from the earlier mentioned proposal that subtelomeric regions are responsible for generation of new diversity, they could also be involved in transcriptional regulation. This phenomenon has earlier been found in the budding yeast *Saccharomyces cerevisiae*, where genes inserted in subtelomeric regions were silenced through a mechanism called telomere position effect (TPE) (Gottschling et al., 1990 and Sandell et al., 1992).

Synteny analysis between the two most completely sequenced genomes of *P.falciparum* and *P.yoelii* has revealed that a high level of synteny exists for all but the subtelomeric regions of chromosomes (Carlton 2002 and 2005). *P.falciparum* chromosome ends can be divided into the telomere and adjacent to it, six telomere associated repetitive elements (TARE 1 to 6; TARE 6 is also called Rep20). Only the telomere itself is conserved enough among different *Plasmodium spp.* to allow hybridisation to the same probe (Figueiredo et al., 2000). Maintenance of telomere ends is performed by telomerase and is important for any rapidly dividing organism to avoid telomere erosion. In addition, chromosome truncations are thought to take place in actively dividing parasites at rates between 5.2 x 10^{-4} and 5.3 x 10^{-3} parasites per generation (Horrocks et al., 2004 (b)). The enzyme telomerase normally function to prevent telomere erosion and to heal chromosome truncations in other eukaryotes. A somewhat larger version of telomerase, *Pf*TERT, exists in *P.falciparum* (Figueiredo et al., 2005 a and b). Active *Pf*TERT is present in all the blood stages and can heal chromosome breakpoints (Bottius et al., 1998 and Sriwilaijareon et al., 2002).

Furthermore, the chromatin structure differs along the chromosome length as it was found that the most distal parts of the chromosomes in *P.falciparum* are devoid of nucleosomes (Figueiredo et al., 2000). The size of the telomere region varies between 960 bp (*P.chabaudi*) and 6700 bp (*P.vivax*), with the remainder of the most commonly studied *Plasmodium spp.* falling within this interval. Also both inter- and intra- chromosomal differences in telomere size can be observed within the same species (Figueiredo et al., 2002). Truncated chromosomes, on which the TARE had been deleted was healed by telomerase exhibited the longest telomere lengths observed (Figueiredo et al., 2002). Interestingly, these healed chromosome ends remained localized at the nuclear periphery but were delocalised from the telomere clusters observed for intact chromosomes. If the truncation had occurred close to a functional promoter of a gene, transcription of telomeric DNA was observed (Figueiredo et al., 2002). In addition, spontaneous deletions of a *var* gene and TARE led to transcriptional activation of an adjacent *var* gene (Horrocks et al., 2004 (b)). This indicates TARE could be involved in transcriptional silencing and are responsible for the clustering of telomeres at the nuclear periphery but not for the nuclear periphery localisation itself. Plasmids carrying the Rep20 element were found to co-localize with terminal chromosome clusters (O'Donnell et al., 2002), suggesting an important role for this element in the telomere clustering. Extensive studies of the relationships between the nuclear architecture and transcriptional regulation have been performed in other organisms and form the theoretical backbone of some recent findings in *P.falciparum*. Therefore, an overview over these findings and their implications are given in the following sections.

1.13 Heterochromatin and gene silencing

Chromosomes in the nucleus are tightly packed to allow for the packaging of 2 meters of DNA into 5 µm for the human genome (Li et al., 2004). Chromatin exists in two states called euchromatin (EC) and heterochromatin (HC). HC is more condensed and is normally replicated later in S phase than EC. In addition, HC is less nuclease sensitive, has a more regular spacing of nucleosomes and is less acetylated (Heinkoff et al., 2001). One of the hallmarks of sequences in HC is the high level of repeat elements, and it has been proposed that evolutionarily, HC has formed in an attempt to inactivate parasitic DNA elements originating from retro transposable elements and viruses (Heinkoff et al., 2001). HC has been widely associated with gene silencing of the yeast mating type loci and it was found that specific changes in the chromosomal architecture around these genes were associated with their activity (as reviewed by Rusche et al., 2003 and Weiss et al., 1998). In the fruit fly, *Drosophila melanogaster*, co-expressed genes were found to be localised in chromosomal regions with a less condensed structure in certain developmental stages (Kalmykova et al., 2005). In the pathogen *Candida glabrate,* subtelomeric HC formation and gene silencing of surface encoding genes has also been observed (De Las-Penas et al., 2003). Most of the silenced genes described above are located in subtelomeric regions, however in mammalian females, one of the somatic X chromosomes is silenced by the formation of highly condensed DNA. It has been speculated that the condensed structure of HC prevents recruitment of the transcriptional machinery, and recent findings suggest that the most upstream gene regulatory factors like the TBP (TATA binding protein) are able to bind to promoters in HC regions, but factors further downstream have a much reduced ability to bind and form the pre-initiation complex (Chen et al., 2005).

1.14 Formation of heterochromatin

HC forms through the interactions between silencing factors and histones. A histone is an octameric protein consisting of two of each of the following subunits: H2A, H2B, H3 and H4. Once formed (Fig. 1.2 a, modified from Grewal et al., 2003), histones interact with DNA and each histone wraps DNA around itself approximately

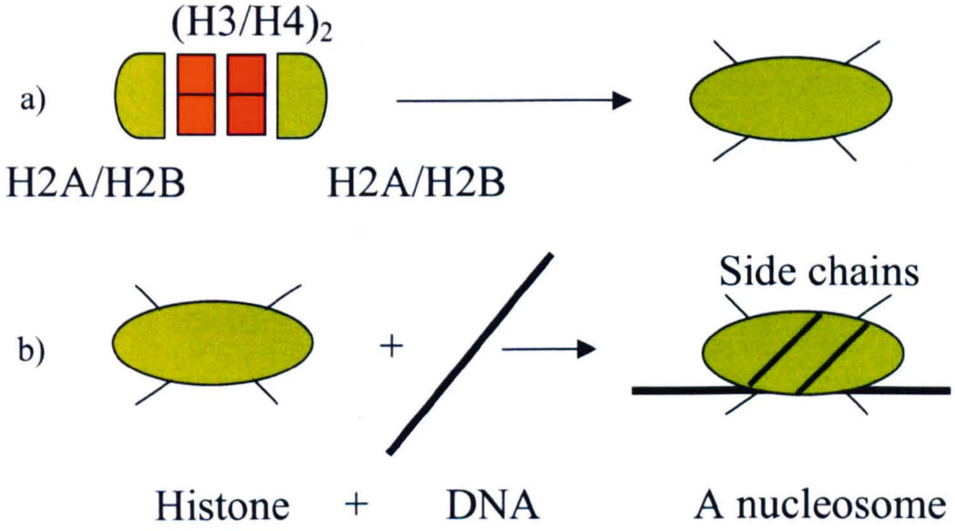
Figure 1.2

Histone assembly

The assembly of a histone (a) and a nucleosome (b).

a) A histone is composed of 8 subunits: 2 x H2A/H2B (shown as one subunit) and 2x H3/H4. These assemble into a histone molecule with histone side chains protruding as shown.

b) The histone molecule interacts with DNA and wraps it around itself approximately two times.


(H3/H4)₂
a)
H2A/H2B
H2A/H2B
Side chains
b)
+
Histone + DNA
A nucleosome

twice (Fig. 1.2 b modified from Grewal et al., 2003). Side chains protruding from the histone are the targets of many posttranslational modifications such as acetylation, deacetylation and methylation. Deacetylation is especially recognized to be one of the mechanisms that form very long stretches of HC.

At the root of the deacetylation machinery are the four silent information regulators (SIR), SIR1 to 4, which are responsible for deacetylation and are found tethered to HC (Gasser et al., 2001 and Rusche et al., 2003). The process initiates when SIR1 is recruited to a silencing site either on its own or through interactions with a silencing factor (Fig. 1.3 a modified from Mozaed et al., 2001 and Grewal et al., 2003). SIR1 then recruits the SIR2/SIR3/SIR4 complex, and SIR2 deacetylates lysines in histone side chains through NAD+ dependent hydrolysis (Shankaranarayana et al., 2003). After the first deacetylation, the SIR complex can bind on its own to deacetylated histone side chains and deacetylate the next histone (Fig. 1.3 b modified from Mozaed et al., 2001 and Grewal et al., 2003). This process leads to repeated rounds of deacetylations and SIR recruitments to histone side chains until a boundary element is reached (Fig 1.3 c modified from Mozaed et al., 2001 and Grewal et al., 2003). Boundary elements are very diverse in sequence (Bell et al., 2001). Some findings indicate that boundary elements recruit histone acetylases and thus establish an EC region (Fourel et al., 1999 and Donze et al., 2001). Other findings suggest that boundary elements physically displace the nucleosomes too far away for the SIR complex to deacetylate it (as reviewed by Rusche et al., 2003).

Another aspect that is observed for HC and gene silencing, which in many cases is formed in subtelomeric regions, is that chromosomes ends are tethered at the nuclear

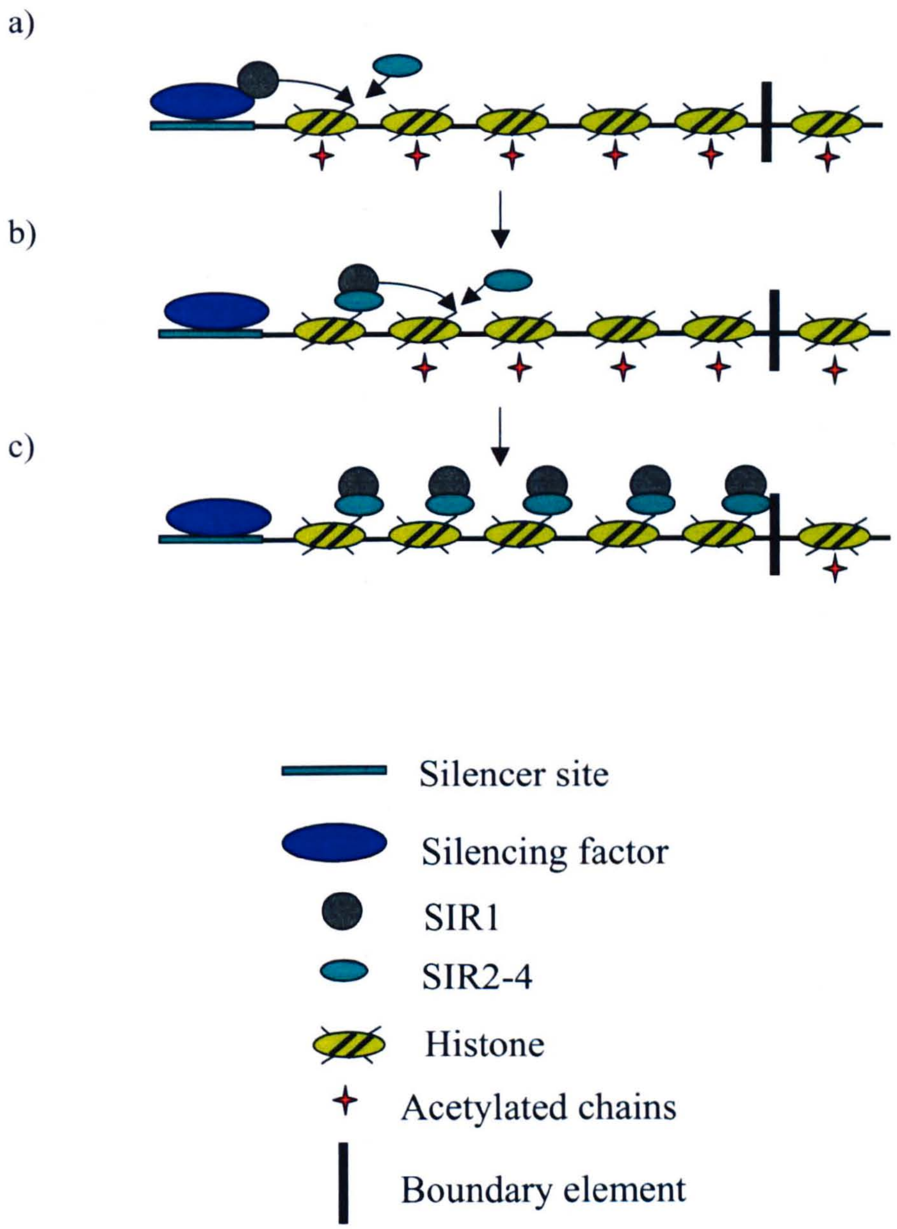
Figure 1.3

Recruitment of Sir proteins to nucleosomes and deacetylation of histone side chains

a) Initially, a silencing factor is recruited to a silencer site, and this in turn recruit Sir1 to the silencing factor. Sir1 then recruits the Sir2/3/4 complex, and Sir2 deacetylates histone side chains.

b) Sir1 recruits another Sir2/3/4 complex to the next nucleosome, leaving the current nucleosome deacetylated and bound to the Sir2/3/4 complex.

c) Eventually this process ends when a boundary element is reached.


a)
b)
c)
Silencer site
Silencing factor
SIR1
SIR2-4
Histone
Acetylated chains
Boundary element

periphery through interactions between proteins at the tip of the nuclear pore complex (NPC) and telomere associated proteins (Feuerbach et al., 2002) (Fig 1.4).

It has been speculated if this tethering was an effect of gene silencing or alternatively if it caused it. It is now clear that the tethering is necessary for maintaining a repressed state as all disruptions that lead to a different localisation of chromosomes also resulted in gene activation and the dispersal of SIR proteins (Feuerbach et al., 2002 and Tham et al., 2001). On the other hand, tethering of chromosomes at the nuclear periphery was not always associated with complete silencing (Tham et al., 2001).

In yeast, a silenced mating type remains stable for up to 10 generations (as reviewed by Rusche et al., 2003) and this type of epigenetic molecular memory is thought to involve either the re-establishment of a particular deacetylation pattern after each cell division or through the random and equal segregations of modified histones, which can then re-establish a particular histone code (as reviewed by Rusche et al., 2003). It seem more likely that the histone code is re-established *de novo* (Fig. 1.5) after cell division than it would be if histones would segregate along with the newly synthesized chromosomes. However, the factors initiating the *de novo* deacetylations could perhaps be segregated during cell division and be able to re-establish the histone code on both parental and daughter strand DNA. It is clear that deacetylation and EC formation is a global way of silencing a large number of genes located primarily in subtelomeric regions and also holds the possibility as a way to maintain molecular memory over several generations.

Figure 1.4

Telomere and nuclear periphery interactions

Schematic of how the telomere ends are thought to interact with the nuclear periphery. In the drawing, a set of proteins (for the sake of simplicity reduced to one) binds to the telomere and interacts with a nuclear pore protein complex. Telomeres kept in this position maintain their coat of Sir proteins and are transcriptionally repressed. If one of the telomere ends detaches from this position, the Sir complexes dissociates and the telomere becomes transcriptionally active.

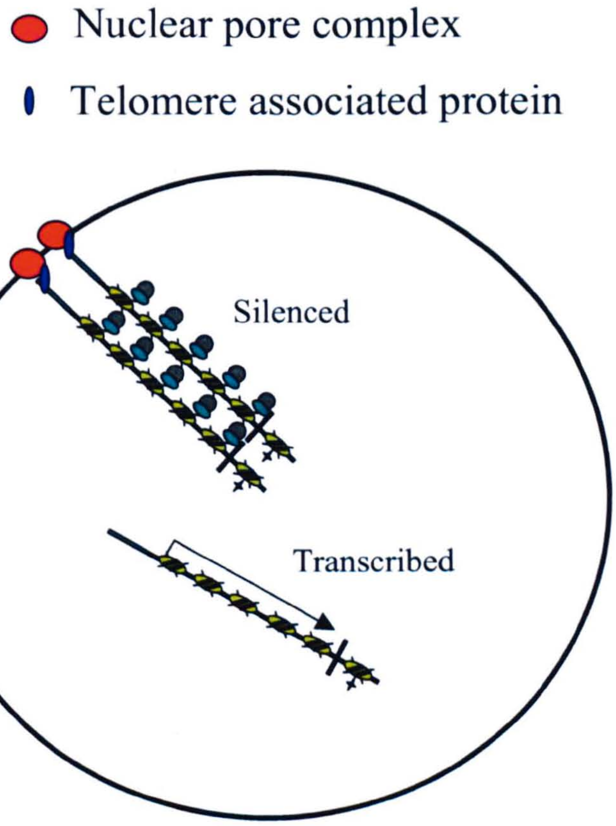
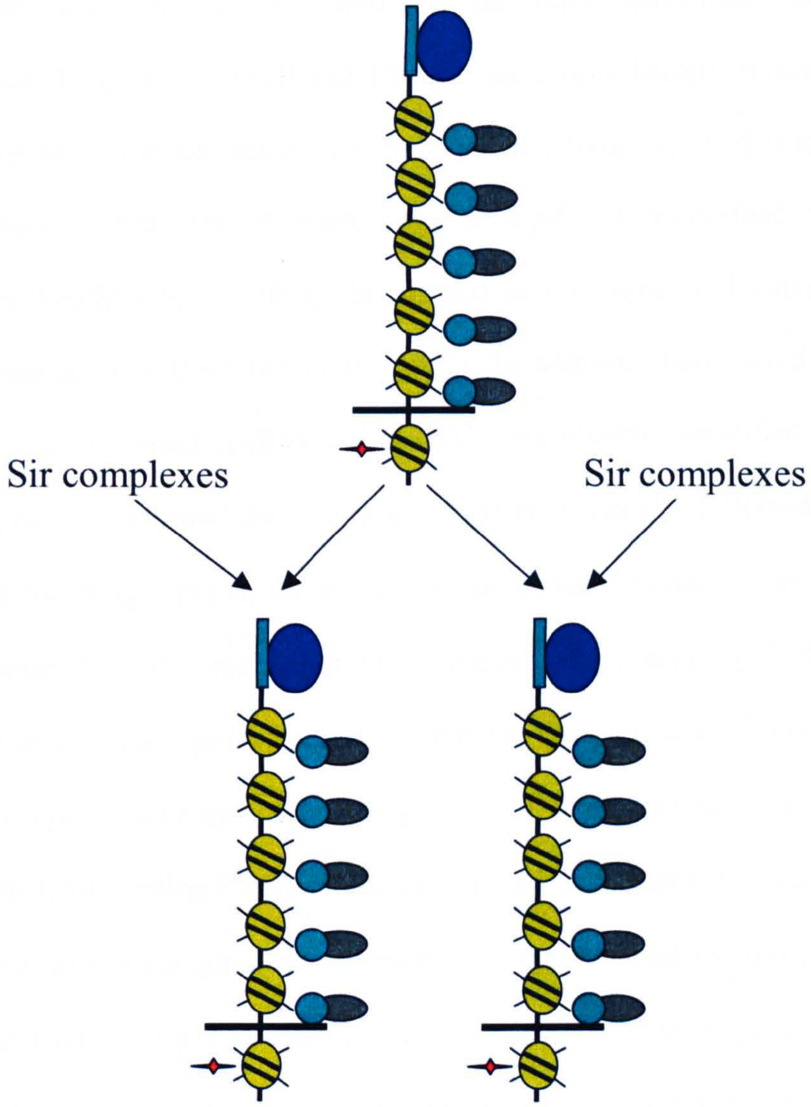

Nuclear pore complex
Telomere associated protein
Silenced
Transcribed

Figure 1.5

Molecular memory

Schematic of how molecular memory can work through Sir proteins. In the top is shown a parental telomere end coated with Sir complexes. The chromosome is then replicated and the histone code is re-established *de novo* on the two newly synthesized chromosome strands.


Sir complexes
Sir complexes

1.15 Regulation of *var* genes

Var genes are preceded by three distinct types of 5′ intergenic regions called upsA, upsB, upsC and (Voss et al., 2000). In the fully sequenced 3D7 strain of *P.falciparum*, 11 upsA, 35 upsB and 13 upsC have been identified (Gardner et al., 2002); upsA and upsB are located in vicinity of the telomeres; upsA is transcribed in the direction towards the telomere whereas upsB is transcribed towards the centromere (Gardner et al., 2002). In contrast to this, upsC is located in internal chromosomal regions (Gardner et al., 2002). In addition, two mixed types of 5′ intergenic regions called upsB/A and upsB/C were recently identified (Lavstsen et al., 2003), and it was found that similarity in ups-type was also reflected in similarity in domain encoding types in the associated *var* genes (Lavstsen et al., 2003). This suggests some kind of homogenisation mechanism of *var* genes in different groups, and could also indicate group specific regulation (Lavstsen et al., 2003). In addition to the ups types shared by several *var* genes, a very distinct *var* gene exist, which encodes the CSA binding Pfemp molecule involved in placental malaria. Constitutive transcription of this *var* gene was observed in different laboratory strains (Kyes et al., 2003) and it was found to contain a unique single copy 5′ region in both laboratory strains and clinical isolates (Vazquez-Macias et al., 2002). *Var* genes are preceded by the subtelomeric upsB and the chromosome internal upsC intergenic regions are transcribed in blood stages and are able to drive a low-level reporter gene expression in episomal transfections (Voss et al., 2000). In addition both types share a conserved 30 bp motif located at different distances from the translational start codon (Voss et al., 2000). Two subtelomeric promoter elements (SPE1 and SPE2) were identified in upsB intergenic regions. SPE1 interacted with nuclear proteins some 18 hours post infection and this correlated exactly with the cessation of transcription (Voss et al., 2003). SPE2 interacted with proteins in a cooperative manner during S-phase of the cell cycle. The central upsC intergenic regions contained a central promoter element (CPE), which also interacted with nuclear proteins at the same time as cessation of transcription occurred. However this was observed 16 to 26 hours post infection (Voss et al., 2003). Internal deletions of SPE1 and CPE followed by episomal transfection led to a small increase in reporter gene activity, but deletions of several other regions where protein binding was not observed did also affect reporter gene activity (Voss et al., 2003).

These proteins are therefore thought to act as repressors of transcription in a coordinated manner. It is interesting to note that subtelomeric and internal *var* gene promoters interact with different proteins and that for central *var* genes this interaction occurs later post infection than for the subtelomeric *var* genes. However internal deletions of the SPE1 and CPE elements only led to a small increase in the reporter gene expression (Voss et al., 2003), and therefore it was thought that epigenetic factors could play an additional role. Factors known from human and yeast to be involved in epigenetic silencing like histone deacetylases (HDAC) and silent information regulators (Sir) have been identified in *P.falciparum* (Scherf et al., 2001 and Joshi et al., 1999). It is therefore possible that these factors are among the nuclear proteins that interact differentially with subtelomeric and central *var* promoter elements.

When *var* gene promoters were removed from their chromosomal context, transcription of a reporter gene was observed regardless of the status of the endogenous promoter (Deitsch et al., 1999). This supported the idea that *var* genes are regulated by epigenetic mechanisms. However, when the *var* intron was included in a transfection construct, silencing of the reporter gene was observed after passage through the S-phase (Deitsch et al., 2001). This showed that the intron is involved in the *var* silencing. However, the necessity for passage through the S-phase before this silencing was observed indicates that chromatin assembly into higher order structures was necessary. In a more recent study, the *var* introns were found to contain a conserved sequence element that resembled a known eukaryotic core promoter, a so-called Inr element (Calderwood et al., 2003). Introns were able to drive reporter gene transcription on their own and one region within the intron was able to silence a *var* promoter (Calderwood et al., 2003). This suggests a role for the intron as an alternative promoter/silencer that competes with the upstream *var* promoter (Calderwood et al., 2003). Recently it was found that elements within the *var* intron were spontaneously deleted when it was placed in front of a drug resistance gene and used in transient transfection (Gannoun-Zaki et al., 2005). This further supported the idea that the intron functions as a silencer. However so far this has only been studied in episomal transfection studies and with internal *var* sequences, so it remains to be seen what the role of the intron is on intact chromosomes. In addition, *var* introns are very heterogeneous in size, ranging from 170 to 1200 bp (Gardner et al., 2002) so it is possible that not every intron is capable of exerting a silencing effect.

Recently, three studies reported how *var* genes were regulated by epigenetic mechanisms. By transfection, Duraisingh et. al. (Duraisingh et al., 2005) integrated the *dihydrofolate reductase* (*dhfr*) drug resistance gene into a subtelomeric region of Chr. 3 via Rep20 TARE. Drug resistant and drug sensitive lines were established and it was determined that drug sensitivity corresponded to inactive *dhfr* associated with closed chromatin, whereas drug resistance, and hence *dhfr* expression, corresponded to a more open structure, as determined by MNase digestion experiments. Furthermore, drug sensitive clones would eventually become drug resistant after expansion. This showed that the silencing is reversible and correlates with alterations in chromatin structure. Deletion of the *P.falciparum*, *sir2* homologue, *Pfsir2,* led to up regulation of especially subtelomerically located *var* genes with upsA type promoters. In addition, *rifin* genes located adjacent to upsA type *var* genes were also up regulated. This showed that the histone acetylase, *PfSir2* is involved in silencing certain groups of subtelomerically located *var* genes. Through the use of 2 colour FISH, they also established that active genes were repositioned to the same regions of the nuclear periphery. This study clearly showed that genes in subtelomeric regions are either active or repressed, and that repressed genes can be derepressed. Furthermore this repression was associated with a more compact chromatin structure and functional Sir2 was necessary for maintaining the repressed state for a number of *var* and *rifin* genes. Activated genes were found to localise to specific regions at the nuclear periphery.

In another study by Freitas-Junior and colleagues (Freitas-Junior et al., 2005), the role of *Pf*Sir2 was further investigated. The authors found that the *Pf*Sir2 protein localized to one pole of the nucleus. By combining telomeric FISH and immunolocalization, it was found that *Pf*Sir2 co-localized with the telomeric clusters. The compactness of chromatin along chromosomes was investigated by using FISH and several probes with known distances between them. This revealed that the chromatin at the telomeric clusters is much more compact than in internal regions of chromosomes. The localisation of acetylated histones and telomeric clusters were found to be mutually exclusive and interestingly Sir2 was found to be associated with the *var*2CSA gene, which is located 55 KB from the telomere. Using Chromatin-Immuno-Precipitation (ChIP) with anti-Sir2 and anti-acetylated Histone 4 antibodies, it was found that the *var*2CSA gene was precipitated with the anti-Sir antibody when it was inactive, but with the anti-acetylated histone 4 antibody when it was active. This *var* gene was flanked by a unique upsE type promoter and was located in subtelomeric regions some 55 KB from the telomere. In contrast, *var* genes in more central regions of chromosomes could only be precipitated with the anti-acetylated H4 antibody when active, but not the Sir2 antibody when they were inactive. This suggests that the subtelomerically located *var* genes are silenced by Sir2 whereas the approximately 35% of the *var* repertoire located in more central chromosomal regions are regulated by a Sir2-independent mechanism.

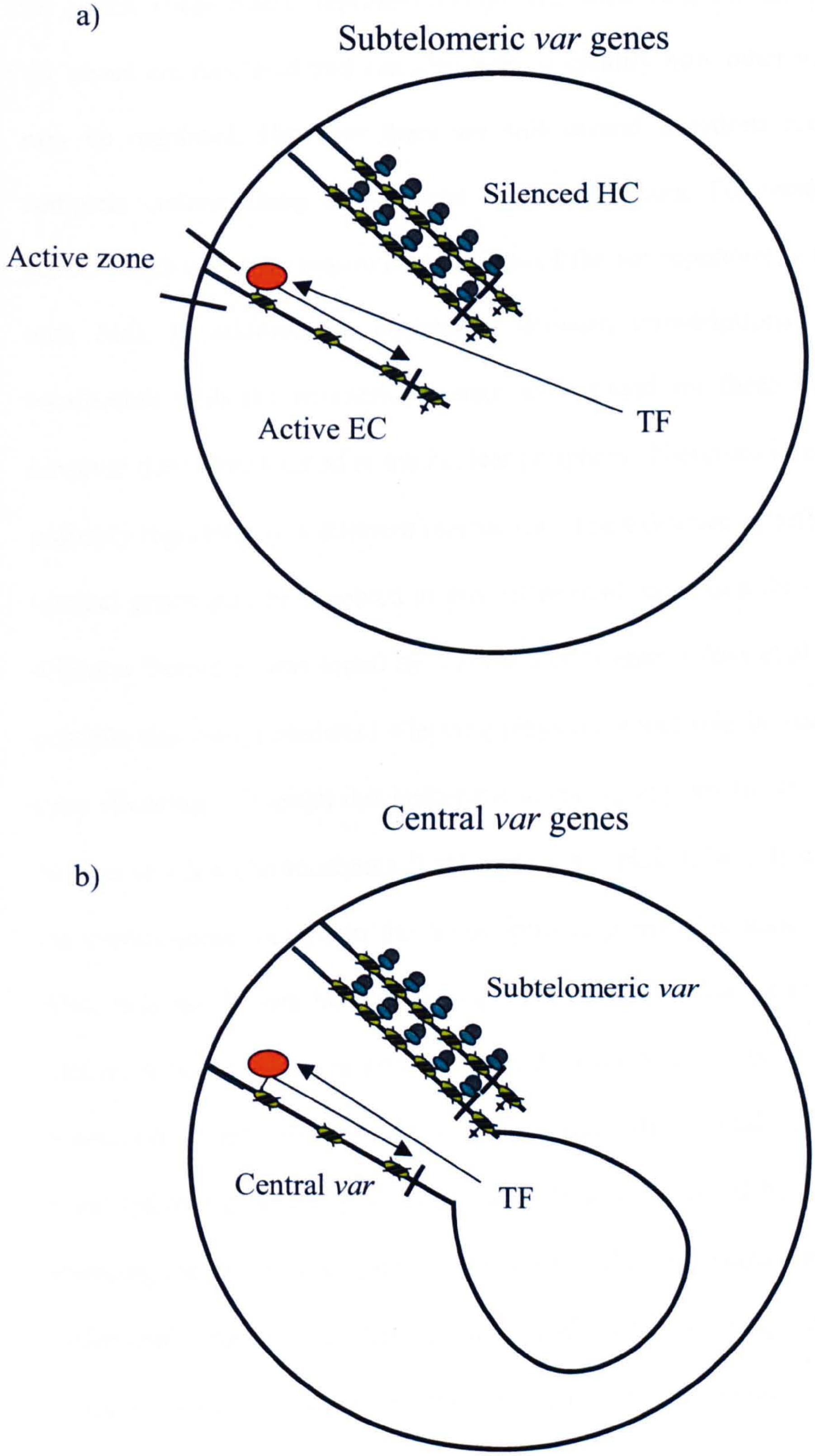
In the third study by Ralph and colleagues (Ralph et al., 2005), the relationship between *var* gene activation and sub nuclear localization was investigated. In this study, the nucleus was divided into three zones (A, B and C) where A represented the outermost zone. No difference in the localisation in the three zones could be observed for active and inactive *var* genes, however internal *var* genes, located some 500 Kb away from the telomere, were observed at the nuclear periphery and this suggest that internal *var* genes loops out to the telomeres. However, active subtelomeric *var* genes were not associated with the clusters of inactive *var* genes, and by examining the nuclear ultrastructure by electronmicroscopy it was found that in all cells studied, one region of less condensed material was located at the nuclear periphery. In contrast, internal *var* genes were randomly associated with the telomeric clusters regardless of whether the *var* gene was active or inactive. In summary (Fig. 1.6 a and b), the subtelomeric *var* genes (a) are associated with Sir2 protein in clusters at the nuclear periphery when inactive and maintain a HC chromatin structure. When activated, Sir2 is no longer associated with them and the active telomere moves away from the telomeric cluster to a less electron dense region of the nuclear periphery. For the central *var* genes (b), looping out to the nuclear

Figure 1.6

Plasmodium var regulation

a) Proposed regulation of the subtelomeric *var* genes. *Var* genes are kept inactive through the clustering together of telomere ends in a tight chromatin state (HC) bound with Sir protein. One of the telomere ends has dissociated from this cluster and is located in a less condensed active area of the nuclear periphery. It has attained a looser chromatin structure (EC) with a larger spacing of nucleosomes and transcription occurs probably through a hypothetical transcription factor (TF) indicated.

b) Regulation of central *var* genes. The region containing central *var* genes loop out towards the nuclear periphery. *Var* genes in this region are not bound by Sir proteins regardless of if they are active or not and neither are their histones deacetylated when inactive. In addition, the centromeric *var* genes can be transcribed regardless of their localization at the nuclear periphery.


a)
Subtelomeric *var* genes
Silenced HC
Active zone
Active EC
TF
Central *var* genes
b)
Subtelomeric *var*
Central *var*
TF

periphery occurs, but the transcriptional competence is not determined by co-localization with the telomere cluster. In addition, no Sir2 is bound to inactive central *var* genes. These studies represent a major step forward in our understanding of how *var* genes are regulated and can also help to identify how other multigene families may be regulated. However there are still several questions remaining before a complete understanding of *var* gene regulation occurs. For instance, internal *var* genes, which comprise approximately 35% of the *var* repertoire, were not associated with Sir2. In addition no correlation between transcriptional activity and co-localisation with the telomeric clusters were found for these internal *var* genes, however they were located at the nuclear periphery. Therefore internal *var* genes are probably regulated by a different mechanism. The existence of different *var* UTRs in internal genes may be involved in this differential regulation through recruitment of different factors as was found by Voss and colleagues (Voss et al., 2003). It is also possible that intron-mediated silencing plays a distinct role in internally located *var* gene silencing. Although this epigenetic silencing appears to affect *var* genes on all but one or a few chromosomes, it still does not explain how only a single *var* gene on the chromosome located in the transcriptional permissive zone becomes activated. Also, it is not known how switching from one active *var* gene to another occurs. Earlier, it was mentioned how epigenetic memory could be transferred from one generation to the next through histone codes. This would establish a conserved transcriptional pattern, however if the histone code could be “swapped” between heterologous chromosomes this could be a possible mechanism underlying switching which could explain why there seem to be different switching rates for different *var* genes, if a hierarchy in the preferred transfer of histone codes exist for *var* genes on particular chromosomes. It also needs to be established if the *rifin* and *stevor* genes located adjacent to the subtelomeric *var* genes are subject to the same type of epigenetic silencing/activation.

1.16 Global transcription patterns in *Plasmodium*

Although epigenetic factors might generally be important for the regulation of multigene families, such as that found for the *var* genes, this type of regulation is superimposed on a more fundamental *modus operandii* resembling the classical type of eukaryotic gene regulation. This type of regulation involves specific transcription factors that are able to activate one gene, or a class of genes.

Therefore, it is important to summarize what is known about this type of transcriptional regulation in *Plasmodium.* Given the complexity of the *Plasmodium* lifecycle, there is a strong need for the parasite to regulate precisely the appearances of certain proteins during each of its life cycle stages. In the rodent malaria, *P.berghei*, specific proteins were found to be expressed only in male and female gametocytes respectively (Khan et al., 2005). In addition, transient transfections with intergenic regions of some of these showed that reporter genes were only active in the same stage where the protein was detected (Khan et al., 2005). This showed that the expression of sex-specific proteins was controlled at the level of transcription initiation. In the asexual blood stages of *P.falciparum*, transcriptome analysis revealed transcription of genes involved in transcription in the ring stage, transcription of genes involved in translation in the trophozoite stage and transcription of genes involved in DNA replication and merozoite invasion in the schizont stage (Bozdech et al., 2003). Likewise, another proteomic analysis of 2400 *P.falciparum* proteins showed a correlation between the physiology and the proposed functions of proteins expressed at each stage, and in addition only 6% of the proteins were expressed throughout the different stages (Florens et al., 2002).

The liver stage of the parasites life cycle has received less attention than the bloodstages. However, several transcripts were detected solely in liver stages, and the majority of these had no assigned function (Sacci et al., 2005 and Gruner et al., 2005). Thus, it is clear that *Plasmodium* is able to regulate precisely the expression of proteins involved in the different stages. It is unsurprising that *Plasmodium* regulates expression of proteins necessary for survival and development. What is interesting is how *Plasmodium* mediates this regulation, and also which series of events trigger the development into another life cycle stage. By a limited microarray, 8 regulatory factors that exhibited differential transcription patterns in a gametocyte producer and a non-producer were identified (Gissot et al., 2004). Most notably, 2 kinases exhibited differential transcription profiles, but also transcripts encoding proteins involved in mRNA splicing, a subunit of RNA polymerase II, a putative histone chaperone and transcription factors were differentially transcribed (Gissot et al., 2004). Although this study points to several potentially different mechanisms of regulation (signal, transduction, transcription initiation, histone assembly and mRNA splicing), the large proteomic analysis of *P.berghei* gametocyte specific proteins indicates that the most differentially regulated class of proteins between male and female gametocytes were kinases (Khan et al., 2005). The downstream targets of these kinases might belong to several genes that are not yet recognized, given the high number of hypothetical protein encoding genes in *Plasmodium*. These potential targets could be involved in any type of process from transcriptional initiation to mRNA splicing. Nevertheless, since several of the above mentioned studies were performed on mRNA, this indicates that a large fraction of genes are regulated at the level of transcription initiation in a similar way as numerous studied eukaryotic genes.

1.17 Classical eukaryotic transcription

In the classical model for eukaryotic transcription initiation (Fig. 1.7 a), the 5′ intergenic region before a gene contains several motifs that can recruit proteins involved in transcription (Brown, 2002). At the transcription initiation site, subunits (TFIIA to F) of the RNA polymerase II enzyme assemble. In some cases this can lead to transcription, but typically *cis*-interactions with proteins bound at upstream regulatory regions are needed to start (or repress) transcription. The upstream regions are the master switches, where it is decided whether the gene is transcribed or not.

Upstream regions are often unique for a gene, whereas the region at the transcription initiation site is able to recruit general RNA polymerase components and is therefore termed the core promoter.

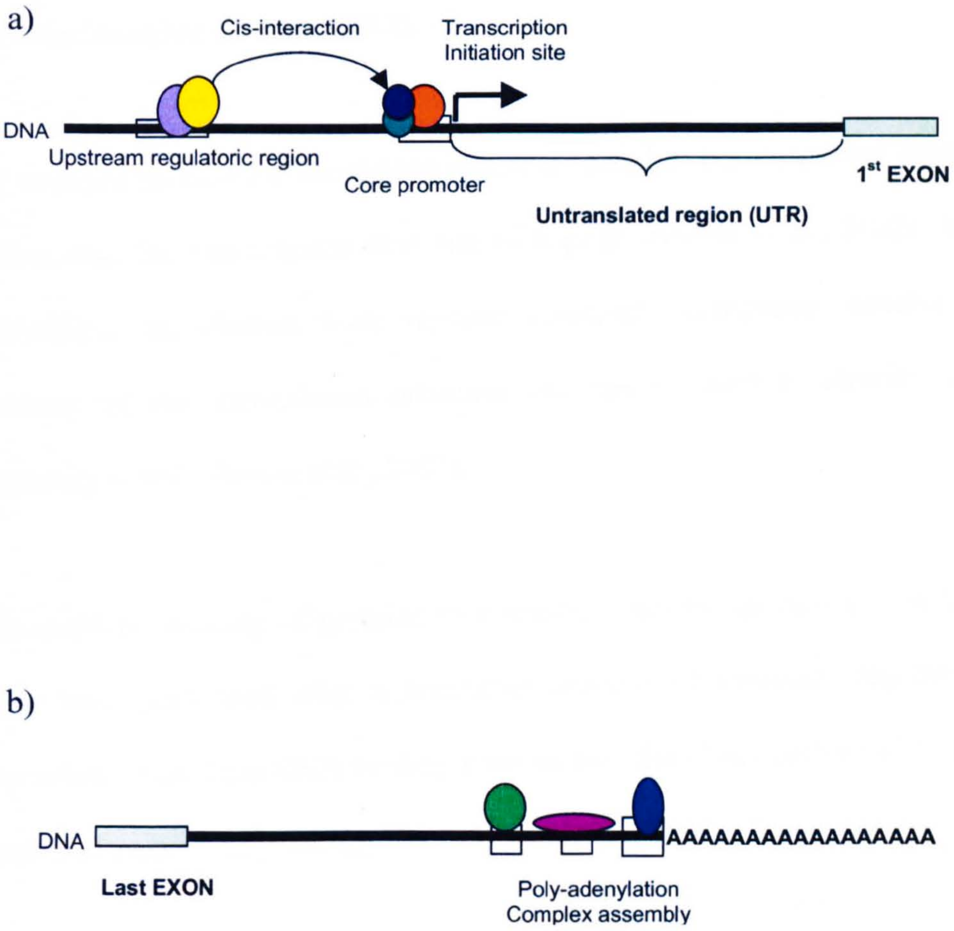
Several well-characterised motifs are known to be responsible for recruiting factors to the core promoter. The most prominent of these are: the TATA box (Consensus: TATAWAW, W: A/T), which recruits the TFIID subunit of RNA polymerase II, the Inr sequence (Consensus: YYCARR, Y: C/T, R: A/G), the CAAT box (Consensus:GGCCAATCT) and the GC box (Consensus: GGGCGG) (Brown, 2002). In this generalized scheme, many eukaryotic genes are specifically regulated by interactions with one particular protein bound to upstream regulatory motifs, which is then able to initiate transcription from a core promoter.

Figure 1.7

Transcriptional regulation

a) Transcription initiation occurs when specific transcription factors binds to the upstream regulatory region. Through cis-interactions these assist in recruiting additional factors to the pre-initiation complex bound further downstream at the core promoter. Through this interaction transcription initiates in the vicinity of the core promoter.

b) Transcription termination occurs at some point after the last coding exon has been transcribed. It involves at least three proteins: one for recognition of a poly-adenylation signal, one for cleavage and one for poly-adenylation of the transcript.


a)
Cis-interaction
Transcription
Initiation site
DNA
Upstream regulatoric region
Core promoter
1st EXON
Untranslated region (UTR)
b)
DNA
AAAAAAAAAAAAAAAAA
Last EXON
Poly-adenylation
Complex assembly

Termination of transcription (Fig. 1.7 b) leads to the addition of a poly-A tail for many eukaryotic mRNAs. The process of transcription termination also involves several factors, and some common elements, like the AAUAAA box, which recruits the cleavage and polyadenylation specificity factor 10-30 nt before the poly A tail has been identified. Two other elements, namely the CA box, which recruits poly (A) polymerase and a GU rich region, which recruits cleavage stimulation factor have also been identified (Brown, 2002).

The classical method for identifying motifs involved in transcriptional regulation is to first map the transcription start site of a gene (Brown et al., 2002). With this information, transfection with reporter constructs containing variable regions upstream of the transcription initiation site can be used to identify important regulatory motifs (Brown et al., 2002).

Alternatively, binding of proteins to interesting motifs can also be used, and has often been performed after a functional analysis of upstream regions. Finally, identification of these DNA binding proteins has often been performed in numerous studies of eukaryotic gene regulation (Brown et al., 2002).

1.18 Motifs important for transcription in *Plasmodium*

In *Plasmodium,* multiple transcription initiation sites have been found for several genes, and this seems to be quite a common phenomenon (Watanabe et al., 2002). The average 5′ UTR lengths of *Plasmodium* transcripts is 346 nt (Watanabe et al., 2002), which is larger than the 103 to 221 NT lengths for transcripts from 7 different taxa (Pesole et al., 2001). There is a well-acknowledged inverse relationship between the GC content of an organism and the UTR lengths (Pesole et al., 2001). A partial explanation for this is that organisms with a high GC content tends to have a higher gene density, which would consequently lead to smaller transcription unit for an average gene (Pesole et al., 2001)

Motifs in the intergenic regions have been found to be important for transcription of a reporter gene in a number of cases. For instance, a G-box conserved between both human and rodent malaria species was found to be important for transcription of the *hsp86* gene (Militello et al., 2004). In another study of the *calmodulin* promoter, a poly (dA)/(dT) tract located upstream of multiple transcription initiation sites was found to interact with nuclear proteins (Polson et al., 2005). A motif located upstream of the *cdp-diacylglycerol synthase* gene was able to enhance transcription by 15-20 fold and did also bind nuclear factors (Osta et al., 2002). In the intergenic regions before the *msp-2* gene, regions with both repressive and activating functions were identified, and it was thought that the repressive regions could play a role in the timing of transcription (Wickham et al., 2003).

1.19 Transcription factors in *Plasmodium*

Although these studies indicate that *Plasmodium* upstream regions contain *cis*-acting elements in promoters that in some cases interacted with nuclear proteins, none of these resemble known eukaryotic *cis*-elements. Overall, there has been a frustrating lack of identified *trans*-factors interacting with *cis*-elements in *Plasmodium*. Of the few *trans*-factors that have been experimentally characterized is the *P.falciparum* *Pf*Myb1, which is a member of the large eukaryotic Myb transcription factor family. Members of this family, which binds to three tandem repeats in DNA, have been reported to be involved in growth control and differentiation. In *P.falciparum*, the *Pfmyb1* gene was identified through bioinformatic methods, and the transcript for this gene was identified in blood stages (Boschet et al., 2004). Differential *Pfmyb1* transcription and DNA binding of the *Pf*Myb1 protein was observed in a gametocyte producer and a non-producer line (Boschet et al., 2004). In another study, the *P.falciparum* TATA binding protein, (*Pf*TBB), was investigated. TBPs are part of the TFIID subunit of RNA polymerase II, and bind directly to promoter elements. It was found that the *Pf*TBP interacted specifically with TATA like elements 81 and 186 bp upstream of the *kahrp* and *gbp-130* transcription initiation sites (Ruvalcaba-Salazar et al., 2005) This distance was larger than the 25-30 bp upstream binding sites normally found for eukaryotes, and in addition there were some sequence variation in the elements it bound to (TATAA and TGTAA compared to the TATAA consensus) (Ruvalcaba-Salazar et al., 2005). In a non-*Plasmodial* study, differences in TATA box sequences have been found to greatly affect the transcriptional activity of a promoter (Hoopes et al., 1998). This was thought to relate to the association/dissociation kinetics of the TBP from different degenerate TATA resembling sequence types, and therefore it is thought that the TBP-TATA interaction is a truly rate limiting step in transcription (Hoopes et al., 1998).

Sequencing of the 3D7 *P.falciparum* strain led to the identification of a surprisingly low number of transcription associated proteins (TAPs) compared to *S.cerevisiae* (Gardner et al., 2002). In a large-scale bioinformatic analysis, only 3% of genes in the malarial genome were found to encode TAPs compared to 10% for seven other crown eukaryotes (current eukaryotes, thought to resemble the first eukaryote in a particular lineage) (Coulson et al., 2004). In contrast to this, all factors comprising RNA polymerase II and various TFII-subunits were clearly identifiable in addition to chromatin remodelling factors such as histone deacetylases (Coulson et al., 2004). This clearly showed that the basal transcriptional machinery and chromatin remodelling factors are conserved in *Plasmodium*. One class of genes encoding a CCCH-type zinc finger domain, involved in regulating mRNA stability, localization and translation, was almost twice as abundant in *Plasmodium* as in the seven other eukaryotic genomes (Coulson et al., 2004). This suggested dominance of post-transcriptional over transcriptional mechanisms as means of gene regulation. However, it cannot be excluded that *Plasmodium* TAPs differ too much from TAPs in normally studied eukaryotes to be identifiable. In a recent study, it was found that three *Apicomplexans* (*Plasmodium, Theileria* and *Cryptosporidium*) contained a lineage-specific expansion of a transcription factor family (AP2) (Balaji et al., 2005). The AP2 family is widely present in plants, and discrete clusters of AP2 genes were found to be expressed in the asexual blood stages of *P.falciparum* (Balaji et al., 2005). Another possibility is that *Plasmodium* uses its transcription factors in a fundamentally different manner. In a study looking at a correlation between conserved motifs in 5′intergenic regions and mRNA levels, it was found that there were more motifs per gene that correlated with transcription in *Plasmodium* than in other organisms (van Noort et al., 2005). The authors propose that this could indicate that coordinated DNA binding of relatively few types of transcription factors determines transcription regulation.

1.20 Evidence for post transcriptional regulation in *Plasmodium*

Studies of how transcripts from individual genes are translated have revealed that in *P.gallinaceum,* the *Pgs28* transcript encoding a protein expressed in the ookinete contains a T rich region in its 3′UTR that act as a strong translational enhancer (Golightly et al., 2000). Further elucidation of this led to the discovery of four regions in the 3′UTR, which were involved in this translational regulation, and these resembled regions in plants and yeast more than metazoans (Cann et al., 2004). In addition, studies of the *SICAvar* genes from *P.knowlesi*, encoding variant surface antigens, have shown that recombination occurs extensively in discrete sequence blocks within the 3'UTR and that these could play important roles for how a particular transcript was processed at the trophozoite stage (Corredor et al., 2004). It has been proposed that *SICAvar* transcripts are subjects to post-transcriptional gene silencing (PTGS) if not destined for translation, and that this could be mediated through sequences in the 3'UTRs (Galinski et al., 2004 and Corredor et al., 2004). However, at current, PTGS of *SICAvar* remain a hypothesis.

The number of genes showing evidence of PTGS has recently been addressed through combined microarray and proteomic analysis, which indicated PTGS for several genes. In the rodent malaria *P.berghei*, nine transcripts were identified in the gametocyte stage, but their protein products were specific to the ookinete stage (Hall et al., 2005). A conserved 47 bp motif was found in the 3' UTRs of six of these transcripts, and this motif was further identified in 29 other genes (Hall et al., 2005). Of these 29 genes, 22 had orthologs in *P.falciparum*, and 16 of these were transcriptionally up regulated in the gametocyte stage, but proteins could only be detected for two of these in the same stage (Hall et al., 2005). Recently, members of the eukaryotic Puf protein family were identified in *P.falciparum* (Fan et al., 2004). Puf proteins are known translational regulators, which they achieve by binding to the nanos-responsive element (NRE) consensus UUGU in mRNA UTRs. Interestingly, in *P.falciparum*, Puf were preferentially expressed in gametocyte stages, and in addition Puf proteins were found to bind to NRE (Fan et al., 2004). The NRE was very abundant in seven of the nine transcripts that showed evidence of posttranscriptional regulation (Hall et al., 2005).

In *P.falciparum*, several proteins involved in glycolysis, merozoite invasion, and the chromatin binding histone H3 protein exhibited a delayed appearance between mRNA and protein (Le Roch et al., 2005). Overall, 319 proteins were detected in stages, where the corresponding mRNA could not be detected. However, many of these were filtered out as they were represented by too few probes on the microarray (Le Roch et al., 2005). Interestingly, especially the gametocyte proteome was rich in proteins with roles in mRNA processing (Le Roch et al., 2005). One problem was though that the microarray appeared to be generally more sensitive than the proteome analysis (Le Roch et al., 2005) and consequently only cases where the protein was detected but not the mRNA could be considered. From these studies it was clear that several genes could be regulated post transcriptionally. Elements located in the 3′UTRs were thought to be important for translational repression in a number of cases, but from the literature there are several other possibilities as to how a delay between mRNA and protein appearance might be regulated.

It must be said here that it is not completely clear at the moment whether *Plasmodium* contains evolutionarily different transcription factors that have so far defied identification. The high number of hypothetical genes with no assigned function could hold surprises. However, from our present view, it is clear that *Plasmodium* is able to exert a highly deterministic control of gene expression despite the apparent absence of numerous transcription factors. It is also clear, from the study of *var* genes, that epigenetic mechanisms are important, and from genome surveys there is a higher than expected abundance of mRNA processing factors. Therefore, there are three known ways of regulating expression of a gene: epigenetic, transcriptional and posttranscriptional. In numerous eukaryotes (mostly metazoans), the overwhelming importance has been on transcriptional regulation, where one gene is regulated by a distinct set of transcription factors. This is highly flexible, as it allows fine-tuning to a wide variety of environmental stimuli. Although *Plasmodium* undergoes several distinctly different stages of development, and is transferred between vertebrate and invertebrate hosts, it spends most of its time in a relatively constant environment with regards to physiological conditions. It might therefore not be so important for *Plasmodium* to fine tune its transcriptional apparatus, which could be reflected in the low number of transcription factors identified. This could mean that epigenetic and post transcriptional mechanisms in combination with relatively few transcription factors could lead to a deterministic expression of genes during the different stages. In contrast to this, it is interesting to note that the genes that seem to be post transcriptionally regulated are transcribed in the gametocytes and are translated in the ookinete. As the transfer between the vertebrate and invertebrate hosts represents the most drastic change in the physiological conditions for the parasite, posttranscriptional activation of already dormant mRNA might be the quickest way to respond to these changing circumstances.

1.21 Regulatory elements in the mRNA

Several features of the UTRs can regulate the rate and timing of translation from an mRNA molecule (Fig 1.8 modified from Mignone et al., 2002). In the 5′UTR, capping and structural elements such as hairpin loops and internal ribosomal entry sites (IRES) can either inhibit or promote the translation efficiency. This is thought to be because several features in the UTR can interfere with the ribosome scanning process. Also, the presence of upstream ATGs (uATGs) and upstream ORFs (uORFs) can reduce the translation rate. The 3′UTR can also regulate the translational efficiency and mRNA stability (Mignone et al., 2002).

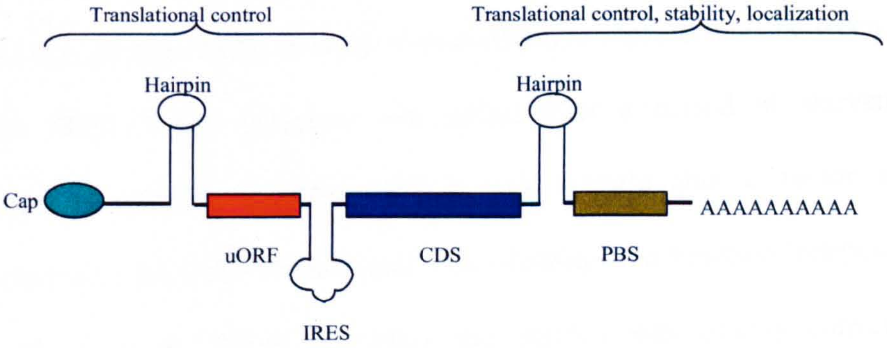
Most eukaryotic mRNAs contain a cap at the 5′ end. The cap interacts with the eukaryotic (translation) initiation factor 4F (eIF4F) complex during translation initiation. One subunit, eIF4A, has helicase activity needed for unwinding of the secondary structure in the mRNA (Mignone et al., 2002). Under stressful conditions, cap dependent translation is repressed as several initiation factors are rendered inactive through phosphorylations by stress kinases (Harding et al., 2000). However, heat shock protein encoding mRNAs are preferentially translated through cap-independent mechanisms when eIF 4 F is inactivated (Barnes et al., 1995 and Joshi-Barve et al., 1992).

Between 15% and 50% of 5′ UTRs from a wide range of organisms contain uORFs, and between 2% and 11% of these contained at least 10 uORFs (Mignone et al., 2002 and Pesole et al., 2001). This is generally thought to decrease the translation efficiency through interference with the ribosome scanning process. However, the sequences flanking the "real" ATG are non-random, and generally fit the Kozak consensus: GCCRCCatgG (R: purine), and therefore the uORFs might be bypassed by the ribosome because it lacks this flanking consensus sequence (Mignone et al., 2002). Secondary structures forming hairpin loops with a stem length of between 65 and 129 nt and with a change in free energy, ΔG below -50 can stall the migration of the 40S ribosome, but this effect can be partially overcome by an increase in the level of eIF4A (Mignone et al., 2002). However if the secondary structures interact with a protein, like the binding of iron regulatory proteins to an iron-responsive element to regulate iron uptake, translation is inhibited for the duration of the interaction (as reviewed by Meijer et al., 2002). There are also some indications, that longer 5′UTRs can affect how efficiently an mRNA molecule could be translated (Yiu et al., 1994)

Figure 1.8

Structural elements in mRNA

A normal mRNA molecule contains a cap at the 5′ end and a poly A tail at the 3′ end. The coding region (CDS) is located between the 5′ and 3′ UTRs, which can contain several structural features such as hairpin loops and IRES. In addition, uORFs in the 5′ UTR can interfere with the ribosome scanning process. In the 3′UTR, certain protein binding sites (PBS) can recruit deadenylation factors or other factors for degradation of the mRNA through nuclease activity. In addition, Puf proteins can bind to such regions and inhibit translation without necessarily degrading the mRNA.


Translational control
Translational control, stability, localization
Hairpin
Hairpin
Cap
uORF
IRES
CDS
PBS
AAAAAAAAAA

Internal ribosomal entry sites (IRES) located in the 5′UTRs function, as the name indicates, as alternative sites from where the translational machinery can load onto the mRNA molecule. IRES were initially discovered in viruses, but a handful of IRES are now thought to be present in various higher eukaryotes (Mignone et al., 2002). Some co-occurrence of uORFs and proposed IRES has been found through bioinformatic analysis of mRNA molecules, and this could suggest that the translational machinery could use IRES to bypass uORFs (Pesole et al., 2001).

In *S.cervisiae*, HAP4 is a protein that is thought to activate transcription after a period of catabolite repression. The 5′UTR of the *hap4* gene both contained a 5′cap and an IRES (Seino et al., 2005). Translation was abolished under normal conditions and this was thought to be because of mutually inhibitory effects between the 5′ cap and the IRES. When galactose was added after a period of starvation, IRES dependent translation occurred, and it was thought that a factor specifically interacted with the IRES element and thus allowing it to function independent of the 5′cap (Seino et al., 2005). Therefore this mRNA was quickly converted into a translatable form when transcription was to resume. This suggests that a shift from cap-dependent to cap-independent translation can occur through the use of IRES. Indeed, various cellular mRNAs exhibits IRES dependent translation when cap-dependent translation is repressed, for instance during stress, apoptosis and in the G2/M phase cell cycle transitions (Meijer et al., 2002). In another study, the 5′UTR of the *myeov* transcript in humans (encoding a protein involved with cell cycle control and misregulated in certain carcinomas) was quite long (445 nt) and contained four uORFs, which on their own severely retarded translation.

However, the 5'UTR was found to contain an IRES that could translate the transcript when mRNA capping was artificially abrogated (Almeida et al., 2005). As a summary, uORFs can inhibit protein synthesis through either stalling or dissociating the ribosome from the mRNA molecule (Fig. 1.9 a), and secondary structures can prevent the ribosome scanning process (Fig 1.9 b).

Finally IRES are thought to recruit the ribosomal subunits to a loading site within the transcript when cap dependent translation or structural features in the 5'UTR prevents initiation or scanning respectively (Fig. 1.9 c).

A recent review raised severe criticism about a number of the published claims for evidence for IRES (Kozak et al., 2005). The common use of bi-cistronic vectors for testing IRES was criticised, as several studies had not checked whether the mRNA was also bi-cistronically transcribed from these vectors. As the IRES elements were introduced between the cistrons, the second cistron (proposed regulated by an IRES) could be mono-cistronic. This criticism was raised because recent studied had shown cryptic promoter activity and splicing between the cistrons. This criticism has been refuted in other publications (Merrick et al., 2005). This author objected that if breakage or splicing occurred between the cistrons, it would leave the second cistron uncapped, and such mRNA species are normally degraded very fast.

Currently, it seems as if there are some problems with the hypothesis of IRES regulation, and as such it should be regarded as a preliminary hypothesis for cellular mRNA species. However, there seems to be more of a consensus regarding the existence of IRES in viral mRNAs.

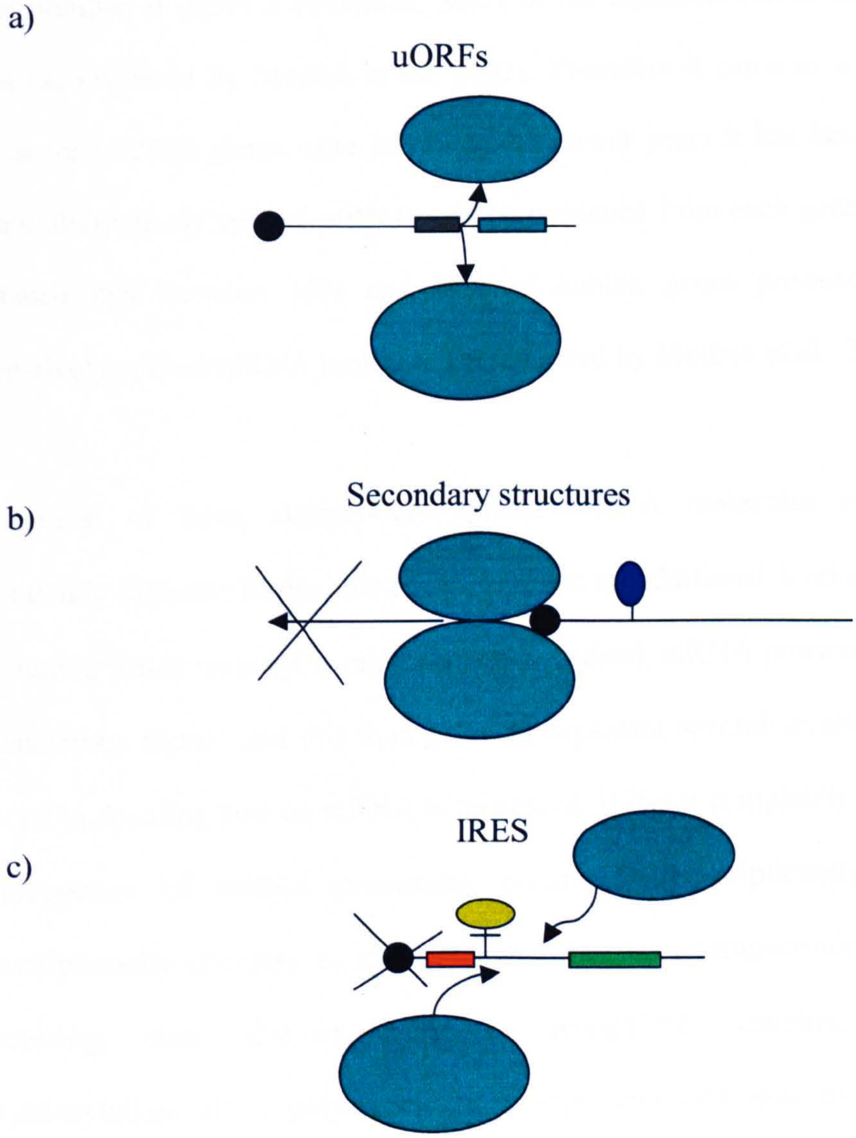
Figure 1.9

Proposed regulatory mechanisms of untranslated regions

a) The presence of uORFs interferes with the scanning process and can lead to either ribosome stalling or dissociation (shown).

b) Secondary structures (as exemplified by the described iron regulatory protein) can interact with proteins under certain conditions and prevent the ribosome scanning process.

c) IRES act by providing the ribosomal subunits with an internal loading site. This is thought that this can occur when cap dependent translation is repressed (through inactivation of eIF4A) or to bypass uORFs (shown in red) located before the coding region (light green).


a)
uORFs
b)
Secondary structures
c)
IRES

1.22 Processing of mRNA

Before the complete sequencing of the human genome there were several estimates of the number of genes it contained. Some of the estimates ran as high as 150,000 genes (as reviewed by Modrek et al., 2002). Therefore it came as a surprise when only around 32000 genes were identified. In recent years it has become clear that several alternatively spliced mRNAs can be produced from each gene, and it is now estimated that between 35% and 59% of human genes produce at least one alternatively spliced mRNA molecule (as reviewed by Modrek et al., 2002).

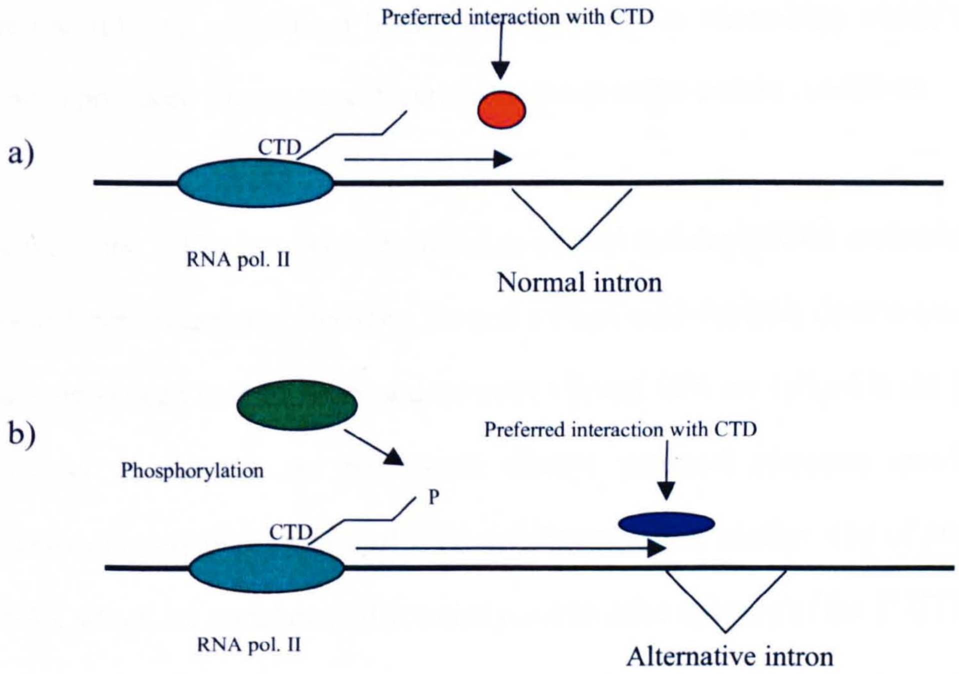
As several of these alternatively spliced mRNA molecules could represent functionally different forms, this could represent an additional level of regulation by fine-tuning which message is produced when. Indeed, mRNA processing is regulated by numerous factors and this thus gives an organism several levers and buttons to operate in deciding how an mRNA is processed. It is not completely clear how large a proportion of mRNA processing occurs co-transcriptionally versus post-transcriptionally (Bentley et al., 2002), but for the co-transcriptionally occurring processing, three distinct steps are recognized: capping, splicing and polyadenylation. RNA polymerase II is especially equipped to cooperate with processing factors through its protruding carboxy terminus domain (CTD) (Bentley et al., 2002). The CTD is composed of between 26 and 52 YSPTSPS tandem repeats in budding yeast and vertebrates respectively (Bentley et al., 2002). The CTD domain is thought to recruit several factors during transcription of a gene and this interaction leads to mRNA capping, splicing, and cleavage/polyadenylation (Bentley et al., 2002 and Fong et al., 2001). It has been proposed that phosphorylations of the CTD when the RNA pol II is still located at the promoter can lead to differential interactions with members of the spliceosome machinery (Fig. 1.10 a and b) and in

Figure 1.10

Proposed roles of RNA polymerase II

a) The CTD protruding from RNA polymerase II normally interacts with one splicing factor (red) leading to the splicing of the same intron in most cases.

b) Sometimes, the CTD is phosphorylated as the transcription machinery is assembled at the promoter, and this phosphorylation leads to the specific interaction with another factor (blue) that leads to alternative splicing of another intron.


Preferred interaction with CTD
a)
CTD
RNA pol. II
Normal intron
b)
Phosphorylation
P
CTD
Preferred interaction with CTD
RNA pol. II
Alternative intron

turn lead to differently spliced mRNA molecules (Bentley et al., 2002). In addition, sequences on the RNA molecule (called exonic splicing enhancers, ESE) can recruit different splicing factors to regions in the vicinity of introns (Brown et al., 2004). Splicing has been associated with at least 145 proteins (58 with no known function) and possibly up to 300 different proteins, which makes it a clear candidate as the most complex molecular mechanism (Zhou et al., 2002 and Jurica et al., 2003). Therefore splicing could be a highly regulated way of controlling which type of protein is produced from a gene in certain stages or under certain conditions.

Splicing in the UTRs has been identified in several cellular mRNA molecules from many different organisms. Between 5% and 35% of mRNAs from diverse taxonomic classes are spliced in the 5′UTR and between 1% and 12% are spliced in the 3′UTRs (Pesole et al., 2001). As this could disrupt potential elements involved in translational regulation, control of UTR splicing could be another way of producing mRNAs, which are translated differentially. Alternative splicing of the 5′ UTR of the *frq* gene in *Neurospora crassa* (involved in circadian clock regulation) was induced by changes in temperature and light and removed an uORF. This lead to translation of another protein isoform encoded by the *frq* gene, and it was thought that the balance between these two isoforms is a part of the environmentally regulated loop controlling circadian rhythm (Colot et al., 2005). Other studies have also shown that alternative splicing of the 5′ UTRs can be induced under various conditions and lead to preferential translation of one of the forms. Prominent examples are the *alas1* gene involved in heme synthesis regulation (Roberts et al., 2005) and the human *nos* gene (Newton et al., 2003).

1.23 Processing and control of mRNA translation in *Plasmodium*

To date, there are only relatively few reports of alternative splicing in *Plasmodium*. Earlier studies of the *stevor* and the *41-3* genes in *P.falciparum* showed that alternative splicing does occur within the coding regions of the gene, and in the case of *stevor* this appeared to happen for distinctively transcribed gametocyte genes (Sutherland et al., 2001 and Knapp et al., 1991). In the most recent study, the *maebl* gene of the rodent malaria *P.yoelii* was investigated.

This gene encodes for three differently sized MAEBL protein isoforms, that were differentially expressed in salivary gland sporozoites and midgut sporozoites (Preiser et al., 2004). In addition, antibodies against MAEBL inhibited sporozoite development in hepatocytes, suggesting it could play a role in hepatocyte invasion (Preiser et al., 2004). At the transcript level, alternative splicing in the 3′ end of the gene produced three different ORFs, two of which lacked the predicted transmembrane domain, and which would therefore not be localized on the surface, (Singh et al., 2004). A temporal difference in the relative appearance of two of the ORFs was observed during oocyte development, and this indicated that alternative splicing was developmentally controlled in this stage (Singh et al., 2004).

Apart from alternative splicing within the coding regions, two studies have reported this to occur in the 5′UTR of the *histidine* gene (Sullivan et al., 1996) and the *b7* gene, thought to encode a protein involved in chromatin assembly (Pace et al., 1998). In the latter, two alternative transcription initiation sites were used in asexual blood stages and gametocytes respectively. When the most upstream transcription initiation site was used in the gametocyte stages, alternative splicing of an intron located 164 bp upstream of the ATG occurred. (Pace et al., 1998).

It has been proposed that the distinct developmentally regulated trafficking of rRNA genes observed in *P.vivax* could be involved in selective translation of different mRNAs during the different developmental stages (Li et al., 1997). However, in the rodent *P.berghei*, knockout of rRNA genes did not inhibit the parasite in undergoing a complete developmental cycle, but it did seem to reduce the growth rate to some degree (van Spaendonk et al., 2001).

Therefore, alternative splicing is occurring in both the coding regions and the UTR of *Plasmodium* genes. The CDS alternative splicing types mentioned above appears to be developmentally regulated to some degree. If this is the case, it could indicate that the parasite uses different forms of these proteins in a stage-dependent manner. Alternatively it could be an extra safeguard mechanism to prevent premature surface localization of proteins. Finally, it could also be a side effect of differential composition/activity of the spliceosome during the different stages that served another, yet undiscovered, purpose. However, splicing appears to be tightly regulated, but it would be interesting to see in a more global analysis if certain stages produce more alternative splicing types than others.

For the 5′ UTR splicing, this appeared to be developmentally regulated as well for one of the types described. 5′UTR splicing could function analogous to some of the mechanisms described earlier, but so far nothing at all is known about 5′ UTR mediated post transcriptional regulation in *Plasmodium*.

1.24 The VIR multigene family

The second most important human malaria species, *P.vivax* is responsible for 70-80 million malaria cases annually and predominates outside Africa (Mendis et al., 2001). However relatively few studies has been performed on this parasite compared to *P. falciparum*. In 2001, sequencing of a subtelomeric YAC clone led to the identification of 32 copies of a three exon multigene family called *vir*. As this number was present on a single subtelomeric YAC clone of 155 Kbp, it was estimated that 600 copies of the *vir* family could exist in *P.vivax* (del Portillo et al., 2001). The 32-vir genes could be separated into five groups, called A to E, based on sequence similarity and size. The VIR proteins contained a single predicted transmembrane region towards the carboxy terminus of the protein.

No similarity to any of the known multigene families in *P.falciparum* could be established at the time (del Portillo et al., 2001). Recently some structural similarity was found to exist between the predicted external VIR protein domain and the newly discovered SURFIN proteins in *P.falciparum* (Winter et al., 2005). The *vir* genes were found to be translated in the asexual blood stages (del Portillo et al., 2001), which made it a new candidate variant antigen, but surprisingly only a couple of *vir* transcripts were detected from the sequencing of a large EST analysis of the *P.vivax* bloodstages (Merino et al., 2003). It was suggested that this finding could either be because the variability in the full *vir* repertoire was larger than initially suspected or because the *vir* transcripts were present at a low level compared to other transcripts (Merino et al., 2003). The known repertoire of *vir* genes were extended through a series of PCR amplifications of genomic DNA from *P.vivax* infected patients, and this led to the discovery of 146 unique *vir* genes (Fernandez et al., 2005).

Through phylogenetic analysis and amino acid similarity it was found that the previously described five groups differed greatly, as no single residue or sequence was shared between them from multiple alignment analysis (Fernandez et al., 2005).

There was also high variability in sizes between the groups: Group A: 90-124 aa, Group B: 193-197 aa, Group C: 185-202 aa, Group D: 173-178 aa. Groups A to C were more heterogeneous than group D, where the variability was mainly observed to occur at the C terminus (Fernandez et al., 2005). Transcription was observed from all groups in the blood extracted from patients, and transcripts from at least two of the groups were detected in individual parasites (Fernandez et al., 2005). Serum from single or multiple infected patients did not differ in the ability to detect 22 expressed VIR tags from the five groups. This suggests that there is no enhanced immune response in multiple infected patients towards specific VIR groups/proteins that leads to clonal VIR expression upon reinfection. However, a generally stronger and more frequent immune response towards VIR proteins in-group A was observed (Fernandez et al., 2005). It has been proposed that the VIR proteins in *P.vivax* have two functions: one set of phylogenetically diverse VIR proteins could function as variant antigens in the circulating blood, while another more conserved (group D, see above) subset could mediate adherence to barrier cells in the spleen (del Portillo et al., 2004).

1.25 VIR homologues in the rodent malarias

Homologues of *vir* were later identified in the three rodent malaria species, *P.yoelii (yir)*, *P.berghei (bir)* and *P.chabaudi (cir)* (Carlton et al., 2002 and Janssen et al., 2002). Intensive investigations of the subtelomeric regions of *P.chabaudi* revealed, in addition to *yir*, 9 other multigene families of variable copy numbers and coding region sizes (Fischer et al., 2003). The *vir/yir/bir/cir* genes have a conserved three-exon structure (Janssen et al., 2002) and are located in subtelomeric regions of chromosomes (Carlton et al., 2002 and del Portillo et al., 2001) and comprises the so far largest known family with 838 annotated *yir* genes in *P.yoelii* (Carlton et al., 2002). The rodent malaria species (*P.yoelii, P.berghei* and *P.chabaudi*) members of this family have been assembled in one joint family; TIGR01590 (HMM ID, Carlton et al., 2002) based on Hidden Markow Models and e value similarities (Carlton et al., 2002, TIGR). The *yir* genes were divided into six different groups based on similarities for predicted ORFs to four *P.yoelii* reference sequences obtained and published earlier as well as one *cir* and one *bir* reference sequence (Janssen et al., 2002 and Fischer et al., 2003)

It has been proposed that the *vir/yir/bir/cir* and the *kir* genes in the monkey malaria *P.knowlesi* forms a supergroup of genes together with the *rif/stevor* family in *P.falciparum*, collectively called *plasmodium interspersed repeats* (*pir)* (Janssen et al., 2004). As this was mainly based on phylogenetic similarities between introns, this proposed relationship does not *per se* imply a functional relationship.

The *vir* and *cir* genes were found to be translated in the asexual blood stages, and in a recent study iRBC surface localisation was found for the YIR proteins especially in the schizont stage (Cunningham et al., 2005). In a large proteomic analysis it was found that BIR proteins were expressed not only in the asexual blood stages, but also in the mosquito, gametocyte and sporozoite stage parasites (Hall et al., 2005). Administration of antibodies against the MSP-8 protein in the lethal *P.yoelii* XL strain led not only to a decrease in the amount of *msp*-8, but also to an up regulation in *msp-1*, *msp-8* and several rhoptry as well as some *yir* transcripts in breakthrough parasites (Shi et al., 2005). This could either indicate that certain YIR proteins are necessary for merozoite invasion or it could simply mean that somehow transcription between the other genes and these *yirs* are linked.

Upon reinfection of immunocompetent mice, changes in the transcription profile for a limited set of *yir* genes were reproducibly observed, and this change was not observed in immuno-deficient mice, where the transcription profile remained unaltered (Cunningham et al., 2005). This demonstrated that at the transcriptional level, *yir* genes are clonally varied when the parasite is exposed to the product of immunological memory. It could also, as mentioned above, be a transcriptional bystander effect of the immune system targeting a highly antigenic epitope. This contrasts to the findings of del Portillo et. al., which said that parasites in multiple infected patients did not clonally express different VIR proteins, compared to first time infected patients. Cross-reactive antibodies to several VIR proteins cannot be excluded especially because the extent of the genomic repertoire is not known. In addition, it was presumed that identical VIR proteins were expressed in previous infections, where no data were obtained.

One of the most interesting recent findings was that the BIR proteins in *P.berghei* were expressed almost completely mutually exclusive in different developmental stages, with 91% of the proteins detected only in a single stage (Hall et al., 2005). This could indicate stage specific functions of distinct sets of BIR proteins, and could mean that this family was involved in several biological processes. As the protein is the result of both transcription and translation, it is not clear exactly at which level this regulation could occur.

1.26 Aims of the project

The overall aim of this project was to identify how *yir* genes are regulated transcriptionally. To do this, I initially set out to investigate how many *yir* genes were transcribed in a population of parasites and in individual iRBCs. This would both give an idea of the extent of the potential repertoire present in the asexual blood stages during an infection, which is an important question from both a molecular and immunological viewpoint. The single cell studies would enable us to assess the extent of transcriptional control over the 838-yir genes.

In addition, mapping of the transcription initiation site was performed, as this would yield important information of where to look for potential regulatory elements. These were then to be assessed further through transfection studies.

It became quickly apparent that with 838 genes it was necessary to have a framework to which experimental findings could be related and also as a starting point for asking questions about function. Therefore, I undertook phylogenetic characterisation, not only of the *yir* coding regions but also of the 5′ and 3′ intergenic regions. Additionally, the gene structure was verified, and it was found that common misannotations had occurred regarding the automated predictions on intron/exon boundaries.

Therefore, this study is divided into the following parts:

- Verification of gene structure and assessment of the gene predictions (Chapter 3)
- Phylogenetic characterisation of the *yir* genes (Chapter 4)
- Transcription in a blood stage population (Chapter 5)
- Transcription in single iRBCs (Chapter 6)
- Identification of transcription start and stop sites and 5′ and 3′ phylogenies (Chapter 7)
- Transfections studies of a putative *yir* regulatory element (Chapter 8)
- Characterisation of an unexpected variety of splice variants (Chapter 9)

Chapter II

Materials and Methods

2.1 Buffers

2.1.1 Giemsa Buffer

0.15M NaCl

0.2mM KH_2PO_4

0.8mM K_2HPO_4

Adjust to pH 7.28

2.1.2 Krebs Solution

6.67 g	NaCl
0.34 g	KCl
0.28 g	$MgSO_4$
222 ml	Buffer (3g Na_2HPO_4, 4.36ml NH_4Cl 217.9ml distilled H_2O, pH 7.4)
2 g	Glucose

2.1.3 Phosphate buffered saline (PBS)

0.1 M	NaCl
80 mM	Na_2HPO_4
20 mM	NaH_2PO_4

Adjust to pH 7.5

2.1.4 TBE Buffer, 10x

100 mM	Tris-HCl
100 mM	Sodium borate
5 mM	EDTA

Adjust to pH 8.0

2.1.5 RNA transfer buffer

7.5 mM NaOH

2.1.6 SSC, 2x

0.3 M NaCl

0.33 M $Na_3C_6H_6O_7$

Adjust to pH 7.0

2.1.7 DNA extraction buffer

50mM Tris

100mM EDTA,

200mM NaCl, pH9

1% SDS

2.1.8 FACS Buffer

PBS (Gibco), pH7.2

1% BSA or Ovalbumin

5 mM EDTA

0.01% NaN_3

2.2 Parasites and experimental animals

2.2.1 Parasites

Plasmodium yoelii yoelii was isolated from the African Thicket Rat, *Thammnomys rutilans* found in the Central African Republic. The non-lethal parasite *Plasmodium* *yoelii yoelii* 17X line A was originally supplied by Prof. D. Walliker (WHO Registry of Standard Malaria Parasites, University of Edinburgh). At the beginning of this PhD, a reference stabilate was generated by passaging an old stabilate, no more than 5 passages away from the original stabilate, through inbred laboratory mice and cloned *in vivo*. Stabilates were kept in liquid nitrogen or at –70°C. During this study, the reference stabilate was used in all studies reported in this thesis.

2.2.2 Mice

Adult female BALB/c mice and wt. female BALB/c mice, or with a targeted disruption of the RAG2 gene (RAG 2-/-) (Shinkai et al., 1992) were bred in the NIMR under SPF conditions, and maintained on sterile bedding, food and water for infection studies.

2.2.3 Parasite infection

Mice were infected intraperitoneally with $5x10^7 – 5x10^8$ parasitised erythrocytes.

2.2.4 Giemsa staining and parasite count

A thin blood film was dried on microscope slides for 30 min, fixed with Methanol for 2 seconds, air-dried and stained with 20% Giemsa staining solution (2.1.1) for 30 minutes. The slides were rinsed under running water and air-dried.

The percentage of parasitaemia was calculated by the following formula:

(No. of infected erythrocytes/ Total number of erythrocytes) x 100%

At high parasitaemia, mice were sacrificed, and blood was collected with 15U Heparin and used in DNA or RNA extraction procedures.

2.3 DNA and RNA manipulations and analysis.

2.3.1 Preparation of parasite genomic DNA

Blood from *Plasmodium yoelii yoelii* 17X infected mice was washed twice in ice-cold phosphate buffered saline (2.1.3) and centrifugated for 5 minutes at 400 x g, at 4°C. The pelleted cells were resuspended in ice-cold PBS (30 mL PBS per mL pellet) and passaged through a Plasmodipur filter (Euro-Diagnostica, The Netherlands) according to manufacturer's instructions, to remove mouse leukocytes. The cells were washed as above and treated with Proteinase K (1mg/ml) in 500□l DNA extraction buffer (2.1.7) overnight at 50°C. The protein was extracted by adding equal volume of Phenol:Chloroform:isoamylalcohol to the digested material and gentle agitation on a rocking platform for 1-2 hours. The aqueous phase was separated from the organic phase by centrifugation for 10 min at 12.000 x g. The aqueous phase containing the DNA was incubated with sodium-acetate and isopropanol to a final concentration of 100mM and 50% respectively. The DNA was collected by centrifugation for 10min at 12,000 x g, air dried and resuspended in nuclease free H_2O (Invitrogen).

2.3.2 Preparation of linearized DNA.

In some cases, linearized plasmid or excised DNA fragments were isolated from the *Pbtubα*-II (See 2.9.1) and the TA pCRII sequencing vector. DNA was purified by resuspension in 1:1 v/v phenol/chloroform/isoamyl alcohol (pH 8.0) and 1:10 v/v 2M Sodium Acetaete. The mixture was centrifugated at 8.5 x g for 15 minutes and the supernatant was carefully removed. To the supernatant, 2 x volumes of 100%

ethanol was added and the sample was left at room temperature for 1 hour, followed by centrifugation at 13000 x g for 30 minutes. The pellet was washed in 70% ethanol followed by a 5 minutes spin at 13000 x g. Ethanol was carefully removed and residual ethanol was allowed to evaporate. The pellet was resuspended in nuclease free H_20 and stored at –20 °C

2.3.3 Preparation of parasite RNA

For RNA extraction, mouse leukocytes were removed as described above (2.3.1). The blood was resuspended in 5x volumes Trizol (Life Technologies) and stored at –70°C until extraction. RNA was isolated by the TRIzol method (Chomczynki et al., 1987). Whole blood, stored in Trizol at –70 °C, was hand warmed, vortexed and resuspended until no clumps were visible. The packed cell volume was resuspended in 10 times 37 °C pre-warmed Trizol v/v followed by the addition 1/5 v/v chloroform, and the tubes were vortexed at full speed for 30 seconds. Then the samples were spun at 7500 x g for 30 minutes. The clear top-supernatant, containing the RNA was removed while carefully avoiding the interphase layer containing DNA and proteins. Equal volume of 100% Isopropanol was added and the sample was stored at 4 °C overnight. The samples were then centrifugated at 13000 x g for 30 minutes and the supernatant was carefully removed. The pellet was washed in 70% ethanol followed by a 5 minutes spin at 13000 x g. Ethanol was carefully removed and residual ethanol was allowed to evaporate. The pellet was resuspended in nuclease free H_20 or formamide and stored at –70 °C. The integrity of the RNA was then checked on a 1.1% Agarose gel containing Guanidinium Thiocyanate.

2.3.4 DNase treatment of parasite RNA

7 µL RNA sample, 1 µL 0.1 M DTT, 1 µL Dnase buffer, 0.5 µL Poly I and 0.2 µL Rnasine was mixed with 0.5 µL Dnase (100 U/µL, GIBCO) and incubated at 37 °C for 1 hour. 30 µL protein denaturing solution (Bio Rad) et al., 20 µL Phenol (pH 4.7) and 20 µL Chloroform was added together with 8 µL 2 M Sodium Acetate. The sample was centrifugated at 7500 x g for 30 minutes and the supernatant was carefully transferred to a fresh tube and equal amount of 100% Isopropanol was added. The mix was kept at 4 °C overnight and treated as described in the RNA extraction protocol.

2.3.5 Gel electrophoresis of RNA

Total RNA from *P.yoelii* (17X A) was analysed on 1% agarose gels (Multi purpose agarose, Roche) containing 5 mM guanidine thiocyanate and 1 µL ethidium bromide per 50 mL of 1xTBE (2.1.3), using trays and combs specially designated for RNA gel electrophoresis. RNA was denatured in formamide for 10 minutes at 65 °C before adding RNA loading buffer and loading it onto the gel. Gel electrophoresis was performed at 65 volts and continued until a good separation could be observed in an UV illuminator.

2.3.6 Gel electrophoresis of DNA

Total DNA from *P.yoelii* (17X A) was analysed on either 1% agarose gels (Multi purpose agarose, Roche) or 3% Metaphor agarose gels (Bio-Rad). Gels were made as described above (2.3.5), except that no guanidine thiocyanate was added. DNA samples were resuspended in DNA loading buffer and loaded onto the gel and electrophoresis was performed at 70 volts for 1% agarose gels and around 90 volts for 3% metaphor agarose gels.

2.3.7 Reverse transcription

RNA was reverse transcribed using random hexamers and Superscript II reverse transcriptase (Invitrogen) according to manufacturer's instructions.

2.4 Stage separation of parasites and micromanipulation

2.4.1 Stage separation of parasites

Different stages of intraerythrocytic parasites (rings (R), trophozoites (Tz), schizonts (Sz)) were separated by density gradient centrifugation using Nycodenz, (Nycoprep, Axis-Shield, Oslo) first preparing a 27.6% Nycodenz working solution (NWS) (in 20mM Tris HCl pH7.4 (Sigma) and then diluting this in RPMI 1640 (AS041-91762 Gibco/Invitrogen, UK) as described below to form the gradients. Schizonts were purified directly on cushions of 50% NWS in medium or, along with trophozoites and rings, on discontinuous gradients of (~2ml) 50, 60, 70 and 80% NWS. Different parasite stages sedimented to each interface and the relative numbers of cells by stage in each fraction was determined from Giemsa's stained thin blood films.

2.4.2 Micromanipulation

Stage separated Schizonts, Trophozoites or Rings were incubated for 30 minutes with a 1:100 dilution of Krebs solution, which was further diluted 3 x 1:10 and incubated with DAPI (which stains parasite DNA in infected erythrocytes). The samples were centrifugated at 3000 x g and the supernatant was removed. 3 washes with Krebs, followed by 3 centrifugations at 3000 x g were performed. Finally 100 μl Krebs solution was added and the different dilutions were loaded on a microwell plate to assess the optimal concentration for picking cells. Single cells were picked under phase-contrast optics, allowing to visualise cells containing DNA, using micropipettes pulled from GC150-F-borosilicate capillary tubes (Clark Electromedical Instruments). Single cells were placed on dry ice immediately after picking and transferred to a pre-heated block at 93°C for 3 minutes to destroy Rnases. After this, the cells were stored at –70 °C.

2.5 Amplification of transcripts from micromanipulated cells

2.5.1 Single cell cDNA synthesis using the SuperSMART kit

The SuperSMART kit (Clontech, BD) was used to amplify cDNA from single cells. This was done by directly adding the single cell lysate into the SuperSMART first strand synthesis reagents and following the manufacturers instructions. It was found that 40 cycles of cDNA amplification was needed to obtain detectable products. The SuperSMART principle is similar to the SMART RACE principle (Fig. 2.1) with the exception that both the modified oligo dT and the SMART II oligo was added simultaneously. These contained the same primer sites, which was incorporated at each end of the transcript. The resulting PCR would then generate a cDNA library containing both 5′ and 3′ ends of a transcript. The manufacturer claims that as little as 50 ng of total RNA can be amplified with this technology. Following the cDNA synthesis, *yir* specific primers were used to PCR amplify *yir* genes.

2.5.2 Single cell aRNA synthesis using the Message Amp kit

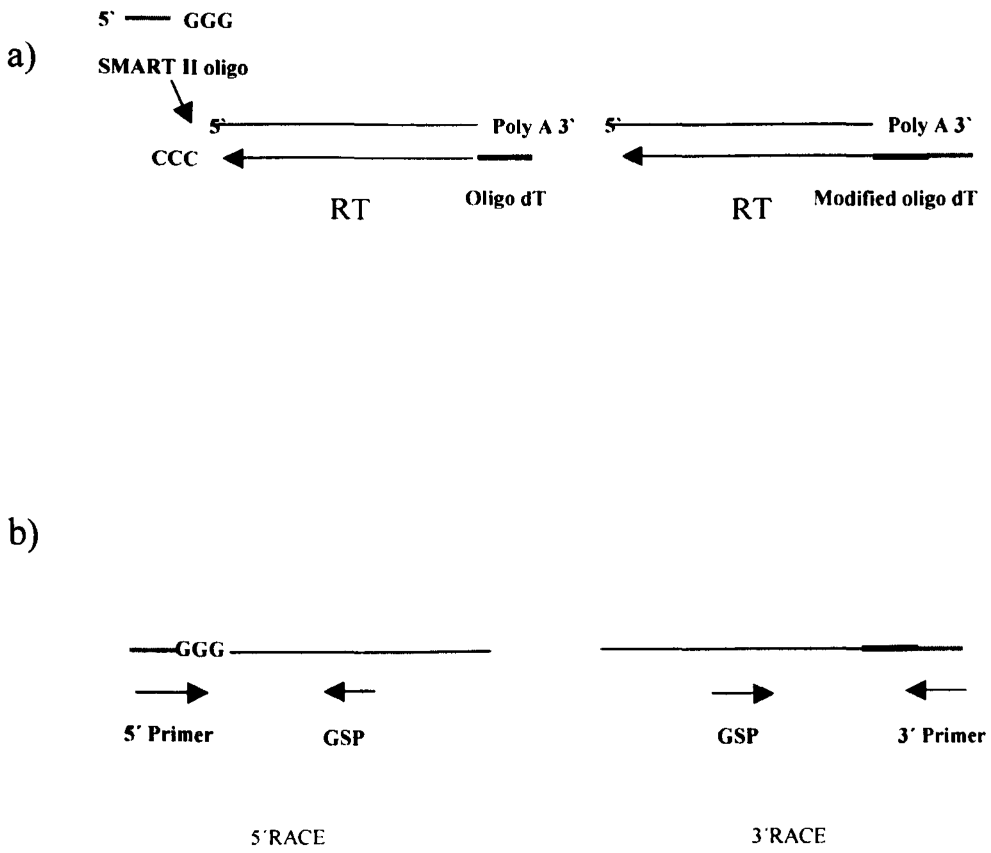
The Message Amp kit (Ambion) was used to amplify aRNA from single cells. This was done by directly adding the single cell lysate into the Message Amp first strand

Figure 2.1

Switching Mechanism At 5' end of the RNA Transcript (SMART)

a) RNA is reverse transcribed (RT) using either a standard Oligo dT primer for 5′RACE or a modified Oligo dT, where a tag is included for 3′RACE. When reverse transcriptase reaches the end of the transcript, it adds three Cytosines. In the 5′ RACE reaction, a SMART II oligo is included, containing three Guanines and a primer sequence. This Oligo attaches to the protruding three Cytosines and template switching incorporates this primer sequence into the 5′end of the cDNA.

b) In both 5′ and 3′ RACE a reverse or forward gene specific primer (GSP) is added to select specific transcripts. Primers recognizing the incorporated tags at either the 5′ or 3′ ends are added, and PCR is performed.


a)
5` GGG
SMART II oligo
5` Poly A 3`
CCC
Oligo dT
RT
5` Poly A 3`
Modified oligo dT
RT
b)
GGG
5′ Primer
GSP
GSP
3′ Primer
5′RACE
3′RACE

synthesis reagents and following the manufacturers instructions. The principle behind this kit (Fig. 2.2) is that aRNA is produced through the incorporation of the T7 promoter site into a double stranded cDNA template. The manufacturer claims that aRNA doubled by a factor of 1000. Here, the resulting aRNA was reverse transcribed using random hexamers (Invitrogen) and PCR was performed with *yir* specific primers.

2.6 Rapid Amplification of cDNA Ends (RACE)

The SMART RACE kit (Clontech BD) was used for RACE according to manufacturers instructions using a forward and reverse exon 3 primer (H) with a Tm of 66.9 °C. The Switching Mechanism At 5' end of the RNA Transcript (SMART) technology (Fig. 2.1) relies on the incorporation of primer tags in the first strand product through template switching at the 5′end. This tag is then used in combination with a primer within the desired gene to amplify regions from the position of this primer to the 5′ of the transcript. The 3′ RACE does not include template switching, but directly incorporates a primer tag with the Oligo dT.

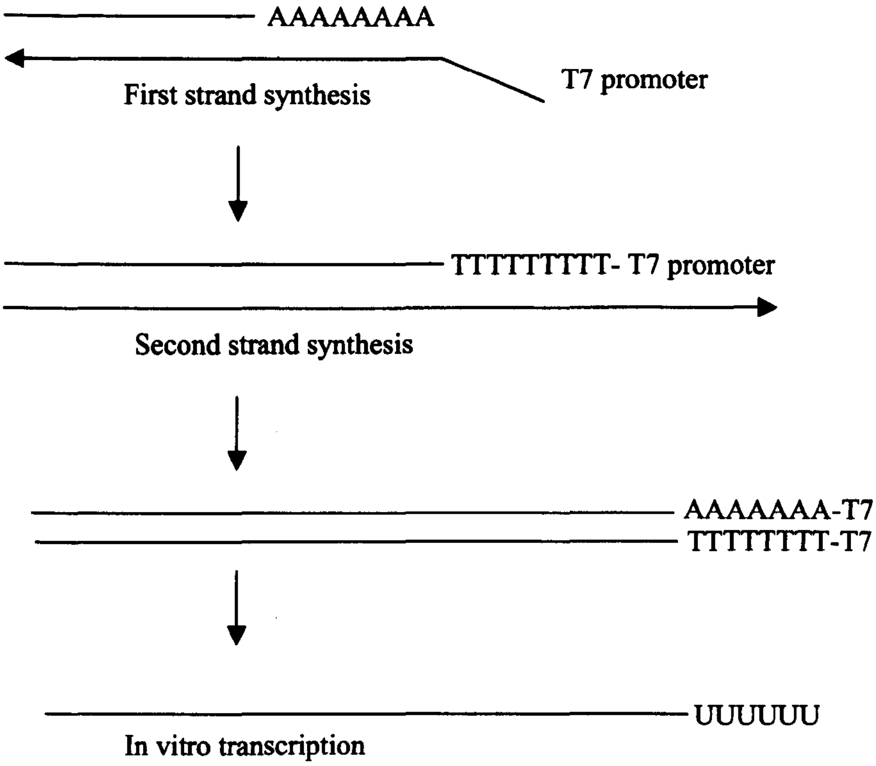
2.7 PCR/RT-PCR and Primers

PCR reactions were performed by adding 2 μL cDNA or DNA into a 48 μL mixture made up with Amplitaq DNA Polymerase or Amplitaq Gold DNA Polymerase (only used once, see Chapter V) (Roche), using the provided PCR and $MgCl_2$ buffers. All PCR reactions were optimised and performed on a temperature gradient Robocycler (Stratagene). For nested or semi-nested PCR reactions, a 1:100 dilution was made from the initial PCR reaction, and 2 μL was transferred to the next PCR reaction.

Figure 2.2

The Message Amp kit principle

The schematic shows the molecular mechanism behind the Message Amp principle for amplifying anti-sense RNA (aRNA). First the RNA is reverse transcribed into cDNA with an oligo dT primer containing a T7 promoter tag. After first strand synthesis, DNA polymerase is added and a second complementary strand is synthesized. The T7 promoter is only functional on a double stranded template. Products are purified by supplied columns, and in vitro transcription is performed with T7 polymerase and rNTPs. After this, the cDNA is degraded with DNase I and the aRNA products are column purified. In this application, the aRNA was reverse transcribed with random hexamers and SuperScript II reverse transcriptase, and PCR was performed with *yir* primers. In this kit, the –RT control was generated by splitting the cells into two samples, and in one sample all ingredients necessary for reverse transcription was added, whereas all ingredients but reverse transcriptase was added to the other half. These samples were designated +RT and –RT respectively.


AAAAAAAA
First strand synthesis
T7 promoter
TTTTTTTTT- T7 promoter
Second strand synthesis
AAAAAAA-T7
TTTTTTTT-T7
UUUUUU
In vitro transcription

Annealing temperature, $MgCl_2$ concentration and cycle number were optimised individually for each primer set. A generalized PCR reaction occurred as follows:

95 °C at 5 minutes (initial denaturation)

95 °C for 1 minute (denaturation, repeated Y times)

X °C for 50 seconds (annealing, repeated Y times)

68 °C for 50 seconds (elongation, repeated Y times)

72 °C for 7 minutes (final elongation)

All primers were synthesised by Oswel (Southampton, UK). The used primer pairs, and the conditions under which they were used (as optimised) are listed in Table 2.1. Locations and amplicon size of the different sets are shown (Fig. 2.3 and Table 2.2). Primer melting temperatures (Tm) were all calculated at: http://www.basic.northwestern.edu/biotools/oligocalc.html. All primers were optimised empirically by assessing band intensity through DNA gel electrophoresis of products generated by varying $MgCl_2$ concentration, annealing temperature and cycle number.

2.8 Preparation of PCR products and sequencing

2.8.1 Purification and cloning of PCR products

Total PCR products that were to be sequenced were purified using PCR Clean Columns (Qiagen). PCR products cut out from a gel were purified using the Gel Extraction kit (Qiagen). PCR products were cloned into the TA pCRII vector (Invitrogen) according to the manufacturer's instructions.

Table 2.1 Primer sets and PCR conditions.

Set[1]	**Sequence**[2]	**Conditions**[3]
Set A	**F** 5′ SMART primer **R** CGGAAACGATTTCAAAAACAAAAATTAAGAG	As recommended
Set B	**F** CTCTTAATTTTTGTTTTTGAAATCGTTTCCG **R** 3′ SMART primer	As recommended

Set C	**F** CGATAAAATTAATGCTGGATGTTTA
	R GTTTTTGAAATCGTTTCCG
		a=4 mM
		b=55 °C
		c= 40

Set D	**F** CAGTATTTAAGTTTTAGAAG
	R TAAACATCCAGCATTAATTTTATCG
		a=4 mM
		b=55 °C
		c= 40

Set E	**F** ATATGGTTAAGTTATATGTTAAACC
	R GTTTTTGAAATCGTTTCCG
		a=4 mM
		b=51 °C
		c= 40

SetF	**F** CGATAAAATTAATGCTGGATGTTTA
	R CCAAATAACGAATACTTATAAGAAATTC
		a=4 mM
		b=53 °C
		c= 45

Set G1	**F** ATATGGTTAAGTTATATGTTAAACC
	R CCAAATAACGAATACTTATAAGAAATTC
		a=2/4 mM*
		b=50/41 °C*
		c= 35/45*

Set G2	**F** ATGGTTAAGTTACAAATTAAACCAAAA
	R CCAAATAACGAATACTTATAAGAAATTC
		a=2 mM
		b=50 °C
		c= 35

Set G3	**F** ATTATGATATGGTTAAGTTATAAATTAA
	R CCAAATAACGAATACTTATAAGAAATTC
		a=2 mM
		b=46 °C
		c= 35

Set G4	**F** ACATTATCATATGGTTAAGTTATAAACTA
	R CCAAATAACGAATACTTATAAGAAATTC
		a=2 mM
		b=48 °C
		c= 35

Set G5	**F** GCTATTTTATGGTTAAGTTATAAAC
	R CCAAATAACGAATACTTATAAGAAATTC
		a=4 mM
		b=53 °C
		c= 35

Set G6	**F** GATAAGATTAATGCTGGATGTTTATG
	R CCAAATAACGAATACTTATAAGAAATTC
		a=4 mM
		b=53 °C
		c= 35

Set H	**F** CTCCTCTCAAAGTGC
	R CTATCGACGAACTTGATG
		a=2 mM
		b=42 °C
		c= 35

Set I	**F** GCAACTTAATTTGAAAATAC
	R CTATCGACGAACTTGATG
		a=2 mM
		b=47 °C
		c= 35

Set J	**F** GCAATATATCTTATCTCTCTCCTCTTAAAGG
	R CGATAAAATTAATGCTGGATGTTTA
		a=3 mM
		b=54 °C
		c= 35

Set K	**F** AAGCTTATAGATACGGGTTCAGTGC
	R CGATAAAATTAATGCTGGATGTTTA
		a=3 mM
		b=54 °C
		c= 35

1: Primer set. 2: Forward and reverse primer sequences. 3: PCR conditions; MgCl$_2$ concentration (a), annealing temperature (b) number of cycles (c). Conditions marked with an * were used only for single cell RT-PCR.

Figure 2.3

Primer locations

Primer locations were estimated on an average *yir* gene. Including both introns, this *yir* was 1120 bp in size, and without the introns 905 bp in size. For intron 1, the average size was 125 bp and for intron 2 it was 90 bp. An average 5′ intergenic region of 1000 bp was also included. Minus positions indicates 5` intergenic sequences, and position 0 to 3 is the start ATG codon. An intron with an average 115 bp size was present in the 5′intergenic regions for some *yir* genes and was located at position -235 to −120. As this intron was only present in some *yir* genes, it is not shown. Set A and B were used in 5′ and 3′ RACE, and no distance could be calculated beforehand for these.

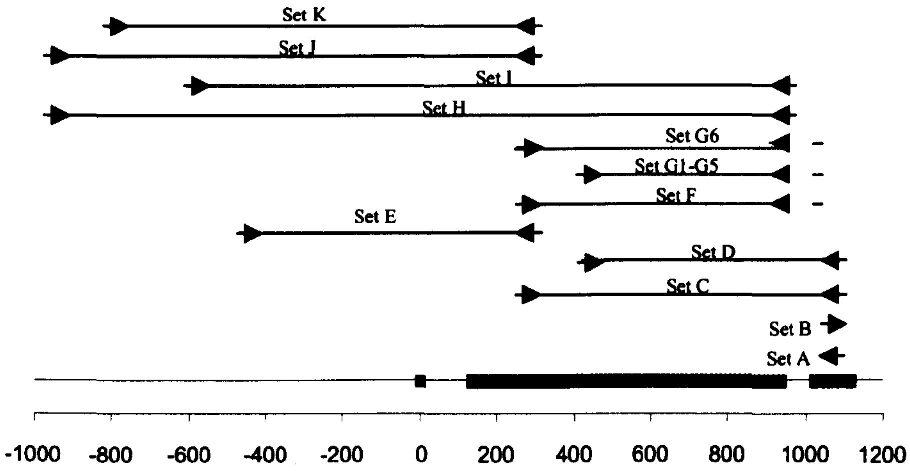

Set K
Set J
Set I
Set H
Set G6
Set G1-G5
Set F
Set E
Set D
Set C
Set B
Set A
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200

Table 2.2 Expected sizes of products amplified with the different primer sets.

Primer set[1]	Size on gDNA[2]	Size on cDNA[3]	With UTR intron[4]
SetA	*ND*	*ND*	*ND*
SetB	*ND*	*ND*	*ND*
SetC	750-850	660-760	*ND*
SetD	600-700	510-610	*ND*
SetE	700-800	575-675	460-560
SetF	*ND*	650-750	*ND*
SetG1-6	*ND*	500-600	*ND*
SetG6	*ND*	650-750	*ND*
SetH	1850-1950	1725-1825	1610-1710
SetI	1550-1650	1425-1525	1310-1410
SetJ	1200-1300	1075-1175	960-1060
SetK	1050-1150	925-1025	810-910

As 90% of all *yir* was between 740 1120 bp in size, the expected product sizes varied with +/- 50 bp for each set. 1: Primer set. 2: Expected size range on genomic DNA (gDNA). 3: Expected size range on cDNA. 4: Expected size range of *yir* genes containing UTR intron.

2.8.2 Transformation

The pCRII constructs were transformed into competent Inv α (Genotype: F′ endA1 recA1 hsdR17 (rk-, mk+) supE44 thi-1 gyrA96 relA1 80lacZM15 lacZYA-argF)U169), TOP10 (Genotype: F- mcrA (mrr-hsdRMS-mcrBC) 80lacZM15 lacX74 recA1 ara139 (ara-leu)7697 galU galK rpsL (StrR) endA1 nupG) or TOP10′ (Genotype: F′{lacIq, Tn10(TetR)} mcrA (mrr-hsdRMS-mcrBC) 80lacZM15 lacX74 recA1 araD139 (ara-leu)7697 galU galK rpsL (StrR) endA1 nupG) cells (from Invitrogen) by heat shock, following the manufacturer's instructions. Transformed cells were grown using the 100 μg/mL ampicilin, following the manufacturer's instructions (Invitrogen). Colonies containing inserts were identified by blue/white screening and grown overnight at 37 °C in 10 mL of LB containing the appropriate antibiotic, following the manufacturer's instructions (Invitrogen).

2.8.3 Small-scale preparation of DNA

Isolation of plasmid DNA was performed using either the Qiagen Miniprep Kit (Qiagen) or the Wizard Miniprep Kit (Promega).

2.8.4 Confirmation of inserts

Inserts in the TA pCRII vector were confirmed by EcoRI digestion for 2 hours followed by DNA gel electrophoresis (2.3.6).

2.8.5 Sequencing

Inserts in the TA pCRII vector was sequenced using M13 Forward and Reverse primers, and in the case of 5′ RACE, forward and reverse versions of the forward primer in Set C was used to obtain complete sequence coverage. The ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing kit (Applied Biosystems) was used according to the manufacturers instruction. Samples were ethanol precipitated before addition of 10 μl MegaBACE™ loading solution (Amersham Biosciences). The MegaBACE 96 capillary machine (Amersham Bioscience) was used for sequencing.

2.8.6 Sequence analysis

Sequence quality was assessed by inspecting the base-tracing capability of the automated sequencing in the *.abi file accompanying each sequence from the automated sequencing. Only regions without ambiguity in base tracing were analysed further. Sequences were analysed by BLASTN against the *P.yoelii* annotated genes at: http://tigrblast.tigr.org/er-blast/index.cgi?project=pya1 and for inspection of the contig location of sequences at: http://www.plasmodb.org/. The sequences were aligned with their corresponding annotated *yir* gene and contig in BioEdit® and manually inspected. The total average BLASTN similarity to database sequences were 89% over a region of 426 bp per average sequence, however this was based on comparison to annotated genes, and thus does not take misannotations of genes into consideration.

2.9 Construction of transfection vectors and transfection experiments

2.9.1 Transfection vectors

Two transfection constructs (Fig. 2.4 a and b) that were previously shown to work successfully in *P.berghei* transfections (Blandine Franke-Fayad et al 2004) (kind gift of Andy Waters, Leiden, The Netherlands) were used in this experiment. The constructs contain *gfp* under control of either the *P.berghei* elongation factor (*Pbef*, Fig. 2.4 a), or the *P.berghei* tubulin α-II promoter (*Pbtub*α-II, Fig. 2.4 b). The *Pbef* promoter was found to drive a strong GFP expression in all stages (Blandine Franke-Fayad et al., 2004), while the *Pbtub*α-II promoter only drove GFP expression in male gametocytes (Blandine Franke-Fayad et al., 2004). The construct could either be used as episomes in transient transfection, or as an integrated unit in the c-ribosomal rna gene unit (c-rna) (Blandine Franke-Fayad, 2004). This integration was targeted through the flanking c-ribosomal rna sequences (*d-ssu-rrna*) fragment in the vector. Integration was not found to affect parasite development (Blandine Franke-Fayad et al 2004). The *Toxoplasma gondii* selectable *dhfr* resistance gene (*tgdhfr-ts*) is under the control of the *pbdhfr-ts* promoter. The two vectors are identical in all aspects, except for the promoters used to drive *gfp*.

2.9.2 Primers used in sequencing of inserts and probe generation

Primers for sequencing and probe generation were designed on the basis of the *Pbtub*α-II vector (see Fig. 2.4 b for schematic and for sequence, supplementary S2.1). In this vector (Fig. 2.4 b and S2.1), the *Pbtub*α-II promoter was flanked by a unique EcoRV restriction site (blunt: GAT/ATC) at position 9238 and a unique BamHI restriction site (Overhang: G/GATCC) at position 10503.

The *gfp* gene was located at position 10509 to 11225. Primers were designed and are listed below with respect to these positions (Fig. 2.4 b and S2.1).

Table 2.3 Primers used in sequencing and probe generation from the *Pbtub*α-II vector.

Primer[1]	Sequence[2]	Position[3]	Application[4]
Seq F	GGACTTGATTTTTAAAATGTTTATA	9129	Sequencing
Seq R	GTCCCAATTCTTGTTGAATTAGATG	10542	Sequencing
GFP F	CCCAGATCATATGAAACAGCATGAC	10730	Probe
GFP R	CTCACACAATGTATACATCATGGCAG	10971	Probe

1: Primer names. 2: Primer sequences, 3: Positions relating to *Pbtub*α-II. 4: Application of primer.

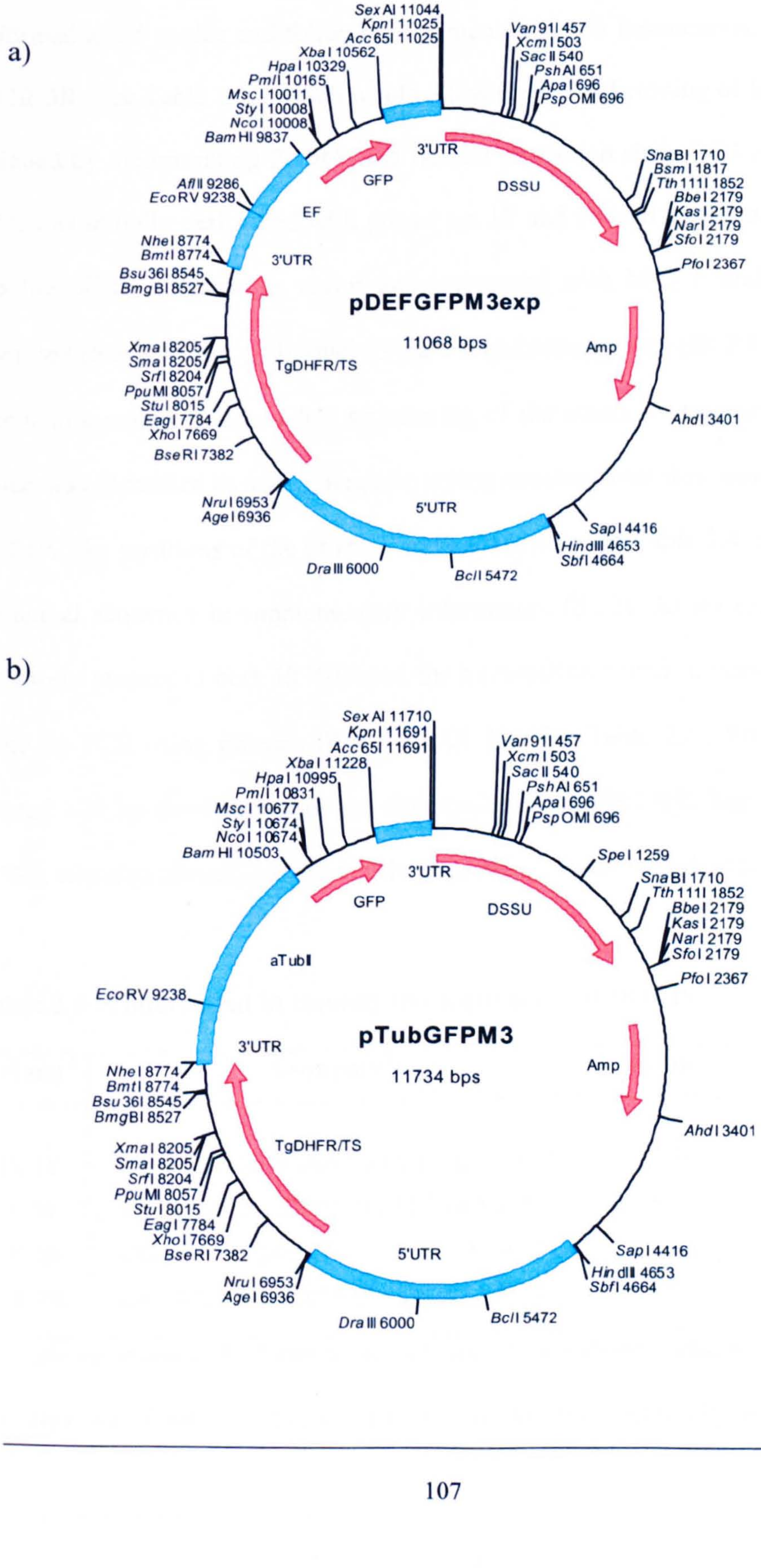
Figure 2.4

Transfection vectors

Schematic of the two transfection vectors. The two vectors are identical in all aspects, except for the promoter used to drive GFP expression. Both contain an *Amp* transcription unit for bacterial growth, a *Toxoplasma gondii* selectable *dhfr* resistance gene (*tgdhfr-ts*) under the control of the *pbdhfr-ts* promoter. In addition the vectors contained flanking c-ribosomal rna sequences (*d-ssu-rrna*), which can be used for stable integration of linearized vector into the c-ribosomal rna gene unit (c-rna). Both figures show unique restriction sites in the vectors.

a) The pDEF GFP M3 exp (called pEF) vector used as a transfection control. This vector contains the *P.berghei* elongation factor promoter (position 9238 to 9837) in front of *gfp*.

b) The pTub GFP M3 (called pTub) vector used as a replacement vector. The pTub II promoter, driving GFP expression, is located at position 9238 to 10503, flanked by the unique Eco RV and Bam HI restriction sites.


a)
pDEFGFPM3exp
11068 bps
3'UTR
GFP
DSSU
EF
3'UTR
Amp
TgDHFR/TS
5'UTR
SexAI 11044
KpnI 11025
Acc65I 11025
XbaI 10562
HpaI 10329
PmlI 10165
MscI 10011
StyI 10008
NcoI 10008
BamHI 9837
AflII 9286
EcoRV 9238
NheI 8774
BmtI 8774
Bsu36I 8545
BmgBI 8527
XmaI 8205
SmaI 8205
SrfI 8204
PpuMI 8057
StuI 8015
EagI 7784
XhoI 7669
BseRI 7382
NruI 6953
AgeI 6936
DraIII 6000
BclI 5472
SbfI 4664
HindIII 4653
SapI 4416
AhdI 3401
PfoI 2367
SfoI 2179
NarI 2179
KasI 2179
BbeI 2179
Tth111I 1852
BsmI 1817
SnaBI 1710
PspOMI 696
ApaI 696
PshAI 651
SacII 540
XcmI 503
Van91I 457
b)
pTubGFPM3
11734 bps
3'UTR
GFP
DSSU
aTubII
3'UTR
Amp
TgDHFR/TS
5'UTR
SexAI 11710
KpnI 11691
Acc65I 11691
XbaI 11228
HpaI 10995
PmlI 10831
MscI 10677
StyI 10674
NcoI 10674
BamHI 10503
EcoRV 9238
NheI 8774
BmtI 8774
Bsu36I 8545
BmgBI 8527
XmaI 8205
SmaI 8205
SrfI 8204
PpuMI 8057
StuI 8015
EagI 7784
XhoI 7669
BseRI 7382
NruI 6953
AgeI 6936
DraIII 6000
BclI 5472
SbfI 4664
HindIII 4653
SapI 4416
AhdI 3401
PfoI 2367
SfoI 2179
NarI 2179
KasI 2179
BbeI 2179
Tth111I 1852
SnaBI 1710
SpeI 1259
PspOMI 696
ApaI 696
PshAI 651
SacII 540
XcmI 503
Van91I 457

2.9.3 Amplification of *yir* intergenic regions for transfection studies

PCR was performed using the High Fidelity kit (HF, Clontech) containing a mixture of normal Taq DNA polymerase and proofreading enzyme. All PCR reactions were performed at 25 cycles and following the manufacturers instructions. Primers IR 1F and IR 3R (See Table 2.4), which would allow directional cloning of fragments, were designed by incorporating EcoRV and BamHI restriction sites. PCR on *P.yoelii* 17X DNA was initially performed with primer set 1F and 1R and the product was cloned into the pCRII sequencing vector and sequenced with M13 F and R primers as described (See 2.8). An additional forward sequencing primer (IR 2 F) was designed within this region to allow full sequencing of the insert. A fragment of 1019 bp, which was identified as a *yir* intergenic region resulted from this, and this was called IR1019. The positions of the primers on IR1019, listed in Table 2.4, are indicated on the actual sequence in supplementary information (S2.2). As no unique restriction sites were present in both IR1019 and the transfection vector, a nested deletion was made by PCR using primers IR 2F and IR 1R (See Table 2.4). Primer IR 2F was located 125 bp downstream of the first nucleotide of IR1019. This product, called IR894, was also cloned into the pCRII-sequencing vector for sub cloning.

Table 2.4 Primers used in cloning and sequencing of IR1019.

Primer[1]	**Sequence**[2]	**Position**[3]	**Application**[4]
IR 1F	C**GATATC**GAGAATGATATACATTAC	0	EcoRV
IR 2F	TATAAATATATTCCTTTTCTATGTTC	633	Sequencing
IR 3F	C**GATATC**GACTAATCAAACATATTT	126	EcoRV
IR 1R	CAAAAATCCCATTATA**GGATCC**ATG	1019	BamHI

1: Primer names. 2: Primer sequences, 3: Positions relating to IR1019. 4: Application of primer. The restriction sites are highlighted in bold.

2.9.4 Generation of replacement vector

In this study, the EcoRV and BamHI restriction sites flanking the pTub II promoter (Fig. 2.4 b) was used to generate a promoterless vector. Digestion of pTub GFP M3 vector was performed for 3 hours at 37 °C with a mixture of 1 U/μL of Eco RV and Bam HI (both 10 U/μL stock, Roche), in buffer B (100 mM Tris-HCl, 1 M NaCl, 50 mM $MgCl_2$, 10 mM 2-Mercaptoethanol, pH 8.0, Roche) supplied with both enzymes and recommended by the manufacturer for double digestion. The empty vector was isolated from the 1265 bp pTub II fragment by gel electrophoresis at half the concentration of ethidium bromide normally used (see 2.3.6). The empty vector was cut out of the gel and was extracted from the gel slice in Dialysis membrane bags (Medicell, International) at 100 Volts for 1 hour in 1xTBE. This was followed by phenol/chloroform extraction (see 2.3.2).

2.9.5 Generation of the negative control vector, pTub (-)

For the generation of a negative control, where no promoter was located in front of *gfp*, the protruding ends in the vector was blunt ended by the large DNA Polymerase I Klenow fragment (Promega) and religated with T4 ligase (Roche). Around 3 μg of the BamHI/EcoRV digested replacement vector (See 2.9.4) was resuspended in a total volume of 20 μL, containing 1 x Klenow buffer (10 mM $MgSO_4$, 0.1 mM DTT, 40 μM of each dNTP et al., 20 μg/mL BSA, Promega) and 3 U Klenow fragment (from a 150 U stock, Promega) for 10 minutes at room temperature. The reaction was stopped by heating to 75 °C for 10 minutes. Ligation was performed by adding 3 U T4 ligase (1 U/μL stock, Roche) and 3 μL T4 ligase buffer (66 mM Tris-HCl, 50 mM $MgCl_2$, 50 mM DTT, Roche) and distilled sterile water to a total volume of 30 μL. Ligation was performed at 22 °C overnight. Ligation mixture was transformed into competent Top10 bacteria and grown under 100 μg/mL ampicilin, and individual clones were purified as described (See 2.8.2 to 2.8.4). EcoRV/BamHI digestion (See 2.9.4) was performed to identify clones where a successful blunt ended religation had occurred.

2.9.6 Generation of experimental transfection vectors

Both IR1019 and IR894 were isolated from the pCRII sequencing vector by 3 hours of EcoRV/BamHI digestion (See 2.9.4). Fragments were isolated as described (See 2.9.3). IR1019 and IR894 fragments were mixed with the empty and linearized replacement vector (See 2.9.4) at various insert to vector ratios (1:1, 1:3, 1:5 and 3:1) in a total of 10 μL containing 1 μL 1 x T4 ligase Buffer (66 mM Tris-HCl, 50 mM $MgCl_2$, 50 mM DTT, Roche), 1 U T4 ligase (1 U/μL stock, Roche). Ligation was performed at 22 °C overnight. Ligation mixture was transformed into competent Top10 bacteria and grown under 100 μg/mL ampicilin, and individual clones were purified as described (section 2.8.3). EcoRV/BamHI digestion (See 2.9.4) was performed to identify clones where a successful blunt ended religation had occurred. Since the 5′ regions before the inserts were repeated twice in the vector, primer Seq F and Seq R (See Table 2.4) were used to amplify the entire insert and this was cloned into the pCRII TA sequencing vector. Sequencing was performed with the standard sequencing primers M13 F and R as described (See 2.8.5). To obtain full sequence coverage of the insert, primer IR 2F (See Table 2.4) was used. PCR was performed with the HF PCR kit (See 2.9.3) to avoid misincorporations of nucleotides. Two clones, which contained correctly inserted IR1019 and IR894, were identified. These are subsequently called pIR1019 and pIR894 respectively.

2.9.7 Preparation of transfection vectors for electroporation

Each of the transfection vectors, pDEF GFP, pIR1019 GFP, pIR894 and pTub (-), a were transformed into Top10 cells and grown overnight at 37 ° C in 500 mL LB containing 100 μg/mL ampicilin. Plasmid DNA was isolated by using the Maxi Prep (Qiagen) method, and a sample was checked by EcoRV/BamHI digestion followed by agarose gel electrophoresis.

2.9.8 Electroporation

P.yoelii 17 X and YM (called transfection 1 and 2 respectively). For each recipient mouse, 100 μl of stage separated schizonts (approximately $1x10^7$ $-1x10^8$) (See section 2.9) were mixed with 40 μg of plasmid DNA in 40 μl PBS and electroporated in a chilled gap cuvette/Gene Pulser system at 25 μF and 800 Volts (Bio-Rad). The cells were immediately injected into recipient mice intravenously.

2.9.9 Pyrimethamine injections

A stock solution of 2.5 mg/mL pyrimethamine was diluted with saline to prepare injection volumes of 0.1 mL, to give a final dose of 19 mg/Kg bodyweight, injected intraperitoneally.

2.9.10 Transfection experiment 1

The pDEF and pTub1019 constructs were electroporated into *P.yoelii* 17X Schizonts, and these were injected into female BALB/c mice (See 2.2.2). Untransfected and mock-transfected *P.yoelii* 17X Schizonts were also injected. Parasites were serially passaged under drug cover through 2 sets of BALB/c mice, each passage lasting 5 days, while daily monitoring parasitaemia. At the end of the second passage, parasitaemia for the various samples were between 2.5 and 8.5% for the pIR1019 and pEF transfected parasites, while it was 27.5% for the mock-transfected parasites. No parasites were detected in untransfected and drug treated parasites. Mice were sacrificed and the blood cells were either stored in TRIzol for RNA extraction of used for FACS analysis.

2.9.11 Transfection experiment 2

The pDEF, pTub1019, pTub894 and pTub (-) constructs were electroporated into *P.yoelii* YM Schizonts, and these were injected into female BALB/c mice (See 2.2.2). Untransfected and mock transfected *P.yoelii* YM Schizonts were also injected. The parasites were passaged under drug cover through three BALB/c mice, while daily monitoring parasitaemia. At the end of the third passage, parasitaemias were between 5.5 and 35% for pIR1019, pIR894, pTub (-) and pEF transfected parasites, while mock-transfected parasites had parasitaemias of 22 to 36.5%. Mice were sacrificed and the blood cells were used for FACS analysis.

2.10 Analysis of reporter gene expression and transcription

2.10.1 FACS

Blood cells were washed twice in 15 mL ice-cold FACS buffer (2.1.8). All incubations and centrifugations were carried out at 4°C. Cells were plated out into V bottomed 96 well plates (NUNC) at 2 x 10^6 cells per well. The cells were centrifuged at 340xg for 3 min to remove the FACS buffer. To visualise DNA containing pRBC, the cells were incubated with 20µl HOECHST33342 (10 µg/ml) for 10 min, both diluted in FACS buffer. HOECHST33342 stained DNA had an excitation wavelength of 350 nm and an emission of 424 nm. The cells were resuspended in 150□l FACS buffer and centrifuged at 340xg for 3 min. This step was repeated twice. The cells were acquired by LSR using CellQuest software (Becton & Dickenson, Oxford, UK). HOECHST33342 was measured in FL-1 and GFP expression was measured at an excitation wavelength of 488 nm and emission at 530 nm (Blandine Franke-Fayad, 2004) in FL-4. In each case, 50000 HOECHST33342 positive events were collected. All measurements were performed in duplicates. Results were analysed in the Flow-Jo programme (Flow-Jo.com).

2.10.2 Extraction and transfer of RNA for Northern analysis

RNA from transfection 1 was isolated (See 2.3.3) and RNA gel electrophoresis was performed with approximately 6 μg of total RNA per sample. Electrophoresed RNA was transferred by capillary action in RNA transfer buffer (See 2.1.5) overnight onto a positively charged nylon membrane (Hybond N+, Amersham) after washing with rotation in RNA transfer buffer for 30 minutes. The membrane was rinsed in 2xSSC and cross-linked at 120 mJoules (Auto-crosslink setting) using a Stratalinker® UV Crosslinker.

2.10.3 Radio-labelling of DNA probe

A 217 bp *gfp* probe was obtained by PCR with primer set GFP F and GFP R (see Table 2.3) in the *Pbtub*α-II vector. Radiolabelling was performed with ~25 ng of probe and ^{32}P-α-ATP using the Prime-It® Random Primer Labelling Kit (Stratagene, La Jolla, USA) according to manufacturer's instructions. The probe was isolated from unincorporated nucleotides by column purification on Sepharose G-50 columns (Amersham) according to manufacturer's instructions. The specific activity of the probe was inspected by Geiger counter readings.

2.10.4 Hybridisation with radio labelled probed

Membranes were pre-hybridised for at least one hour in RNA hybridisation buffer (Invitrogen) prior to probe addition. After probe addition, membranes were hybridised at 65 °C overnight. Membranes were then successively washed at 65 °C in: 2x15 minutes (2xSSC, 0.1% SDS), 1x15 minutes (0.5xSSC, 0.1% SDS) and 1x10 minutes (0.1xSSC, 0.1%SDS). After the final wash, membranes were washed with 2xSSC (no SDS) and exposed to Kodak Biomax MR single emulsion film at -70 °C for various lengths of time before development.

2.11 Phylogenetic analysis

2.11.1Construction of Phylogenetic trees

Phylogenetic trees were constructed in the programmes: Molecular Evolutionary Genetics Analysis 2.1 (MEGA2.1®, Kumar et al., 2001) and Phylogenetic Analysis Using Parsimony 4.0b (PAUP 4.0b®). Guide trees were constructed in ClustalW® and viewed in TreeView®. Annotated genes were obtained from www.tigr.org and, if needed, translated into proteins using the standard genetic code. A list of the contig number and positions for all *yir* genes (Carlton et al., 2002) was used to specify sequence lengths in either direction. The contig number and positions were then used to retrieve 5′or 3′intergenic regions at www.plasmodb.org by using the sequence retrieval tool.

Protein trees

For protein trees, 718 *yir* genes were translated into amino acid sequences after removing all sequences beyond the end of the second exon. These protein-coding sequences were aligned with Clustal W®. Based on this alignment, a Neighbour-

Joined (NJ) and a Minimum Evolution (ME) tree was generated, both bootstrapped with 1000 replicas and using the amino acid Poisson correction parameter. From this, unrooted phylogenetic trees were constructed. Groups were identified visually on the trees, and sequences were colour coded accordingly.

DNA trees

For DNA sequence trees, 718 *yir* genes (exon 1 and exon 2 only) or 306 5′intergenic regions or 292 3′ intergenic regions were aligned individually with Clustal W®. For each alignment, a Neighbour-Joined (NJ) and a Minimum Evolution (ME) tree was generated, both bootstrapped with 1000 replicas and using the Nucleotide Kimura 2 parameter. From this, unrooted phylogenetic trees were constructed. Groups were identified visually on the trees, and sequences were colour coded accordingly.

For Maximum Parsimony (MP) analysis, each ClustalW® alignment (*yir* gene, 5′and 3′ intergenic regions) was saved in the nexus format and imported into PAUP. A MP tree was generated and viewed in TreeView ®.

2.11.2 Comparison of trees

Sequences were compared between the trees by exporting the exact sequence order into excel, where the position on the tree and sequence groups were assigned to each sequence. This allowed for comparison of the exact sequence positions and groups between the different trees.

2.11.3 Evaluation of bootstrap values

For the nucleotide trees constructed in MEGA2.1®, bootstrap consensus trees were constructed by gradually increasing the Cut-off value and visually inspecting the resulting tree.

2.12 Sequence analysis methods

2.12.1 Identification of conserved motifs in intergenic regions

Twenty-four contigs containing two *yir* genes in a tail-to-head orientation were retrieved from www.plasmodb.org. From the graphical display option at this site, it was checked that no predicted gene was located between these *yir* genes. The sequences from the translational stop codon of the most 5′ distal *yir* to the translational start (ATG) codon of the next *yir* gene were isolated by manual inspections. The lengths of the intergenic regions varied from 1978-2835 nt, with an average length of 2403 nt. These sequences were uploaded to http://meme.sdsc.edu/meme/website/meme.html for (MEME) analysis, and the program was set to identify the six most conserved motifs and return the analysis as text. The average location of the six motifs were identified and plotted onto the average 2403 nt intergenic sequence.

2.12.2 Promoter predictions

Promoter and polyadenylation predictions were performed using various programmes at:http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF_links.htm, or: http://www.bioinformaticsonline.org/links/ch_09_t_6.html.

2.12.3 RNA structure predictions

An alignment of 110 *yir* UTRs was used for RNA structure predictions. Structures were predicted at: http://www.genebee.msu.su/services/rna2_reduced.html by using default conditions.

2.12.4 Exonic enhancer predictions

Exonic enhancer predictions were performed at: http://rulai.cshl.edu/tools/ESE/.

Chapter III

Gene structure

3.1 Introduction

Currently, the predicted *yir* gene structure is based on the experimentally verified three-exon structure of two *cir* genes in *P.chabaudi* (Janssen et al., 2002). So far *yir* gene annotation is based on automated predictions and to date no manual curation has yet been performed.

3.1.1 Objectives

Therefore, by RACE and RT-PCR across splice sites (3.2.1) it was tested if *yir* genes follow the three-exon structure deduced from the *cir* gene model. In addition, the quality of the current genome annotation was assessed by closer inspection of the automated gene predictions (3.2.2).

3.2. Results

3.2.1 Determination of *yir* gene structure

The *yir* intron/exon borders were investigated by RACE and two RT-PCR reactions (Fig 3.1 a). For the 5′ RACE (Fig. 3.1 b), a band of approximately 1.7 KB was observed while for the 3′ RACE a smear was observed with a more distinct band around 800 bp. For the PCR across the exon 2/exon 3 boundary (Fig. 3.1 c), a single band of ~ 690-790 bp was seen from cDNA compared to the corresponding 770-880 bp PCR product seen from the gDNA (Fig 3.1 c). This approximate 80-90 bp reduction in size of the dominant band is consistent with the predicted size of intron 2. RT-PCR was also performed with primers across the first intron (shown in Chapter IX of this thesis).

Figure 3.1

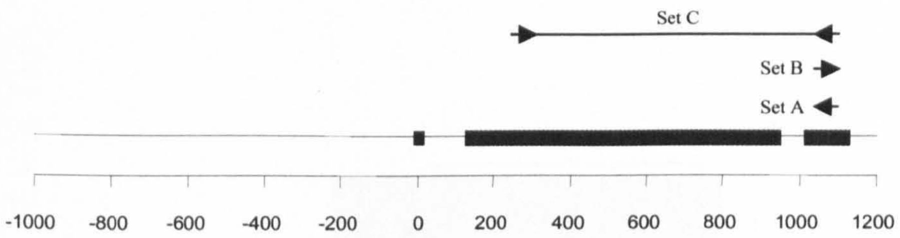
Verification of the gene structure of *yir*

a) Primer location on a *yir* gene. Primer sets A and B were located in exon 3 and used together with the 5′and 3′ RACE primers respectively (see materials and methods). Primer set C would amplify fragments with an average size of 765 bp from gDNA and 675 bp from cDNA. Primer set D would amplify fragments with an average size of 615 bp from gDNA and 525 bp from cDNA.

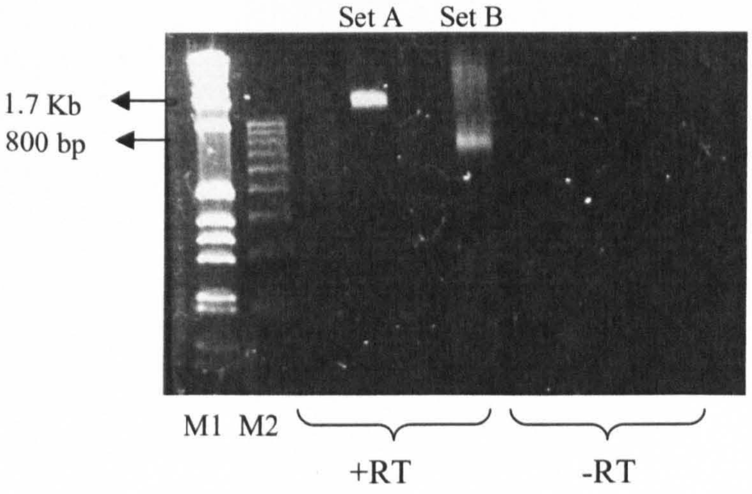
b) 5′ and 3′ RACE was performed with primer sets A and B and according to manufacturers instructions (See materials and methods). Two unlabeled lanes contain attempted 5′ and 3′ RACE with exon 2 primers, which did not work. Both a + and – reverse transcriptase reaction was performed and the absence of products in the –RT lane shows that only cDNA had been amplified. For the 5′RACE with primer set A, a product of 1.7 Kbp can be seen when comparing to the 1 Kb marker (M1), and for the 3′ RACE with primer set B, a band of around 800 bp can be seen when comparing to a 100 bp marker (M1). Both products were cloned and sequenced.

c) PCR with primer set C on cDNA and gDNA and control (water) was performed. PCR on cDNA resulted in a broad band with a center of around 710 bp and on gDNA the center of a similar broad band was around 800 bp. This was as expected from the primers location and it can be seen that there is a shift of around 90 bp in mobility when comparing to the 100 bp marker (M2), which is consistent with the size of the second intron. The cDNA products were cloned and sequenced.

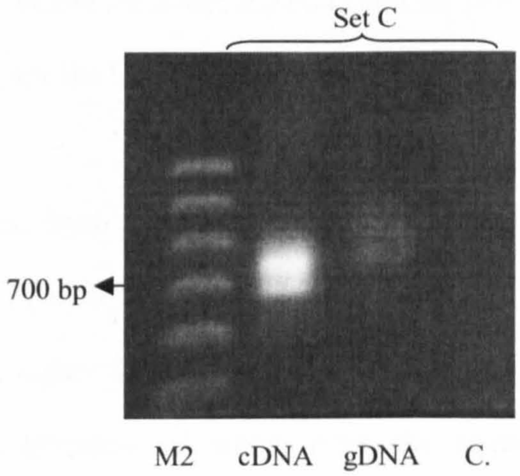
a)


Set C
Set B
Set A
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200

b)


Set A
Set B
1.7 Kb
800 bp
M1
M2
+RT
-RT

c)



However, for now, sequencing of both RACE and RT-PCR products confirmed the expected splicing pattern, consistent with the predicted introns. The cloning and sequencing of 5 different *yir* genes obtained by 5' RACE further confirmed the positions of the first and second intron. Sequence analysis of both RT-PCR and RACE products identified exon 1/exon 2 junctions from 15 different *yir* as well as exon 2/exon 3 junctions from 11 *yir* genes. The junction sites were defined as the nucleotides of an exon before the donor site (GT) and the first nucleotide sequence of an exon after the acceptor site (AG).

The identity of genes, from which the sequences originated were obtained by BLASTN against the *P.yoelii* CDS database at www.http://tigr.org. In all cases, the best match was much higher than the second best match to annotated genes, which allowed a clear identification of which gene the sequence originated from. Interestingly, two of the identified exon 1/exon 2 junctions and three of exon 2/exon 3 junctions did not correspond to the predicted gene model for the annotated genes. Tables 3.1 and 3.2 shows the predicted junctions (TIGR annotation) along with the experimentally validated junctions.

Table 3.1. Sequences of various exon 1/exon 2 splice site junctions obtained by RACE and RT-PCR.

PY name[1]	TIGR annotation[2]	Sequence[3]
PY05822	agt/gtg	agt/gtg
PY00500	cgt/tgt	cgt/tgt
PY03837	gtg/tgt	gtg/tgt
PY02614	cgt/gtc	cgt/gtc
PY06245	gtg/tgt	gtg/tgt
PY07293	gtg/tgt	gtg/tgt
PY05873	gtg/tgt	gtg/tgt
PY04514(*)	gaa/tgt	gaa/tgt
PY00842(*)	gcg/tgt	gcg/tgt
PY03729(*)	gtg/tgt	gtg/tgt
PY04266(*)	tta/tgt	tta/tgt
PY02298(*)	gtg/tgt	gtg/tgt
PY03195(*)	**caa/tgt**	**ata/tgt**
PY04006	**att/ttt**	**atg/ttt**

1: PY locus identifier. 2: TIGR annotation of joined exons. 3: Sequences of joined exons. Transcripts marked with an * are from a 5′ RACE experiment and allowed for simultaneously sequencing of both intron/exon junctions. Bold letters indicate junctions where the sequencing and the TIGR annotation did not correspond.

Table 3.2. Sequences of various exon 2/exon 3 splice site junctions obtained from RACE and RT-PCR.

PY name[1]	TIGR annotation[2]	Sequencing[3]
PY04514(*)	aag/tatt	aag/tatt
PY00842(*)	aag/tatt	aag/tatt
PY03729(*)	aag/tatt	aag/tatt
PY04266(*)	aag/tatt	aag/tatt
PY02298(*)	aag/tatt	aag/tatt
PY07330	aag/tatt	aag/tatt
PY04939	aag/tatt	aag/tatt
PY03195(*)	**aag/ggtt**	**aag/tatt**
PY06131	**aaggtaa**	**aag/tatt**
PY01720	**aaggtaa**	**aag/tatt**
PY07501	**aaggtaa**	**aag/tatt**

1: PY locus identifier. 2: TIGR annotation of joined exons. 3: Sequences of joined exons. Transcripts marked with an * are from a 5′ RACE experiment and allowed for simultaneously sequencing of both intron/exon junctions. Bold letters indicate junctions where the sequencing and the TIGR annotation did not correspond. For these three the TIGR annotation did not predict any splicing, instead it predicted the gene to continue into the second intron.

The RT-PCR and 5'RACE sequence data shown in Table 3.1 and 3.2 verify that the *yir* genes follow the predicted gene model for *P.chabaudi cir* genes. In addition, 5'RACE showed that all *yir* were simultaneously spliced at both junctions.

Incorrect splice site identifications in the TIGR annotation were identified in 14% (2/14) of the sequenced transcripts for the exon1/exon2 splice site and 27% (3/11) for the exon2/exon3 splice site. All five genes could easily be re-annotated manually to agree with the sequencing data. The TIGR annotation was performed through automated gene finding algorithms, and no manual curation has yet been performed (Carlton et al., 2002). These data indicates that there may be a significant number of *yir* genes that are incorrectly annotated. It was important to establish the reliability the TIGR annotation, as this would affect primer design and other analysis.

3.2.2 Evaluation of the TIGR *yir* annotation

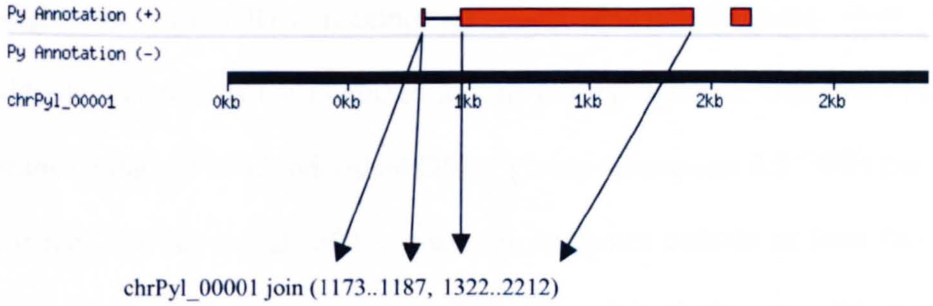
Knowing that the sequenced *yir* genes all conformed to the predicted gene model, it was important to assess the overall accuracy of the automated splice site predictions. To do this, the locations of all the *yir* genes were obtained from a list provided by TIGR. The format of this location was for example for a *yir* gene on contig number 1 (Fig. 3.2): chrPyl_00001 join (1173..1187, 1322..2212). This expression has been used in the automated prediction, and for this example, the predicted annotated *yir* gene consists of two joined open reading frames (ORFs) located at positions 1173 to 1187 and 1322 to 2212 on the contig. Using the "join" function, a proposed *yir* gene was created. In this case (shown), a third ORF was also found in the vicinity of the 3′ end of the annotated gene. By manual inspection this ORF was found to be the third exon, which (in this example) was not included in the original TIGR-annotated gene due to the inclusion of the second intron in exon 2 until a stop codon was eventually reached, and this prevented the third ORF from being joined with the two first ORFs.

For all *yir* genes, the positions of all predicted ORFs were retrieved in the format described above, and split up into each ORF identified. If for instance only the ORF corresponding to the first exon was retrieved (from the example in Fig. 3.2), the argument would have been: chrPyl_00001 1173..1187. The PY locus identifier was also retrieved along with the positional arguments for each *yir* gene in the format:

Figure 3.2

Schematic representation of how an annotated *yir* gene was mapped on a contig

Three ORFs can be seen on the contig, and two of these have been joined to make the annotated *yir* gene. The arrows indicate which positions were joined and the chr_Pyl00001 refers to the contig number. Due to an internal error on the Plasmodb site, the scale shown underneath the ORFs does not match the scale.


Py Annotation (+)
Py Annotation (-)
chrPyl_00001
0kb
0kb
1kb
1kb
2kb
2kb
chrPyl_00001 join (1173..1187, 1322..2212)

PY00001 for the above mentioned example. It can be seen (Fig. 3.2) that the gap between the two joined ORFs is a predicted intron. In order to retrieve the predicted introns used in the annotation, the positions between the end of one ORF and the beginning of the next ORF were used. In this case the format would be: chrPyl_00001 1187..1322.

A total of 122 predicted *yir* genes were found to contain more than two introns, and manual inspection of a number of these revealed that the introns mainly occurred within the largest ORF presumed to correspond to exon 2. These genes were therefore removed from analysis, leaving a total of 716 *yir* genes for further investigations. All ORFs making up these annotated genes were retrieved individually, as well as the predicted introns joining them. In total, the 716 putative *yir* genes contained 1802 individual ORFs, giving on average 2.5 ORFs per *yir* gene. This meant that almost all of the analysed *yir* genes contain at least two putative exons, but it also meant that only 50 % of them contain all three possible exons.

Since the individual ORFs could be for instance both exon 1 and exon 2 for one gene and exon 2 and exon 3 for another, the ORFs were manually selected for the potential exon to which they would correspond. This selection was based on identity to known manual annotations. For instance, ORFs corresponding to the third exon was very easily identifiable because of its high conservation. ORFs corresponding to exon 1 and exon 2 were also both easily identifiable because of the huge size differences between them. After manually sorting ORFs, it was found that there were 637 predicted exon 1 ORFs, 716 predicted exon 2 ORFs and 449 predicted exon 3 ORFs. In addition 637 predicted intron 1 and 449 intron 2 sequences were retrieved.

The automatically predicted ORFs and introns were then manually divided into five separate groups corresponding to 85% of all the predicted *yir* genes. It is known that the GC content of introns (lowest GC content), intergenic regions and coding regions (highest GC content) differs (Carlton et al., 2002) and from the sequencing across exon/intron borders it was found that introns invariably started with GT and ended with AG. Therefore, the overall GC content of presumed introns would be expected to be lower than the retrieved coding regions and these presumed introns would also be expected to start with GT and end with AG.

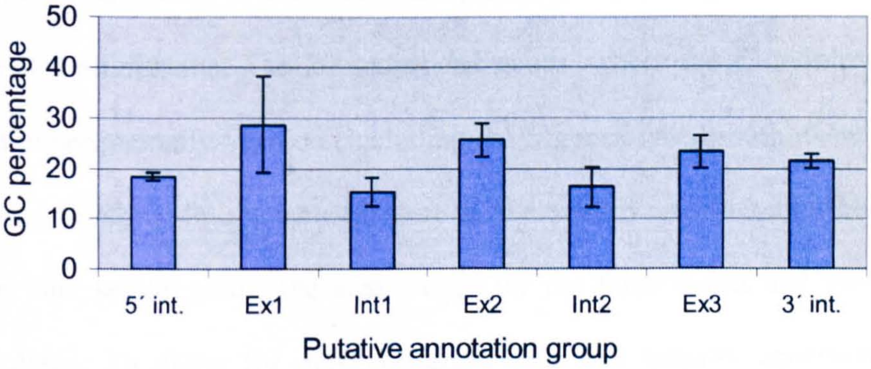
Intergenic regions were included in this comparison as well as they were also expected to have a lower GC content than coding regions. 306 5′ intergenic sequences and 368 3′ intergenic sequences were included. This analysis would give an indication whether the GC parameter is a good indicator of putative exons and introns. The average GC content of each of the five groups and their experimental standard deviations (Fig 3.3) makes it clear that the two putative intron annotation groups have a lower GC content than the putative exon annotation groups and the intergenic regions. This would indicate that the overall success in discrimination between coding regions, introns and intergenic regions based on GC content was quite accurate.

However, it was also clear from the standard deviation for the putative exon 1 sequences that considerable variation exists. It is possible that the first exons contained a subpopulation more resembling the GC content of intergenic regions. Another possibility was that, due to the short sizes of exon 1, random differences in GC content would give a higher variability and thus standard deviation.

Figure 3.3

GC percentages

The calculated average GC percentages for all sequence parts in the five annotation groups. In addition, GC percentages of 306 5′ and 368 3′ intergenic regions are shown. All the parts were retrieved and handled as individual sequences throughout. The bits are indicated as: 5' int (5′ intergenic regions), Ex 1 (exon 1), Int1 (intron 1), Ex2 (exon 2), Int2 (intron 2), Ex3 (exon 3), 3' int (3′ intergenic regions). The standard deviation for each group is also shown on the figure. Dunn`s Multiple Comparison test was performed in Prism to compare the medians between the different annotation parts. This analysis showed that all annotation parts had different median values (P<0.001).


50
40
30
20
10
0
GC percentage
5′ int.
Ex1
Int1
Ex2
Int2
Ex3
3′ int.
Putative annotation group

A second criterion used was to establish whether the putative introns had the GT-AG structure typical of the vast majority of eukaryotic introns. Indeed, all *yir* introns had this structure, showing that the automated algorithms had been highly successful in identifying putative introns using these two criteria.

Having established in principle that the automated program had identified putative coding regions and introns, correct annotations were also evaluated by determining the size distribution of the individual parts, thus avoiding the need for large-scale sequence alignments. The 20 sequenced genes, where the exon/intron borders had been experimentally verified (including the 5 genes where both exon/intron borders were established simultaneously) were used manually annotate a further 35 *yir* genes. From this set of genes, the size ranges for the three exons and two introns were calculated. To allow for more variation than was actually observed within these genes, 10% size variation was allowed except for exon 1 sequences which were invariably either 15 or 18 bp in size. This resulted in the following size ranges being considered a good estimate of correct annotation: **Exon 1**: 15 or 18 bp, **Intron 1**: 100-150 bp, **Exon 2**: 600-1000 bp, **Intron 2**: 80-100 bp**, Exon 3**: 60-100 bp. The following percentages of each annotation part fell within these size ranges: **Exon 1: 54%, Exon 2: 88%, Exon 3: 86%, Intron 1: 60%, Intron 2: 76%.**

3.3 Summary and discussion

In this chapter, the proposed intron/exon boundaries were experimentally verified, and several commonly occurring cases of misannotations identified in automatically predicted *yir* genes.

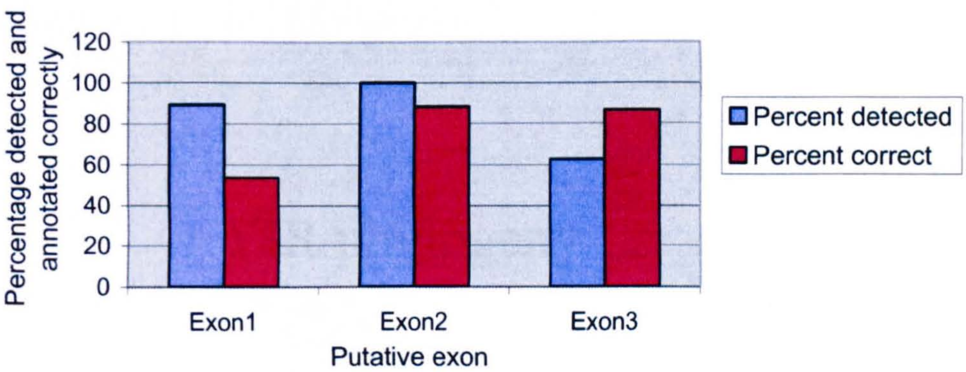
In several cases, the TIGR annotation for the exon1/exon2 and exon2/exon3 splice junctions did not correspond to that found experimentally. From this, it was clear that automated predictions have some limitations. Overall, a high success rate in identifying exon 2 sequences was found. In addition, reasonable accuracy in defining introns and coding regions based on differences in GC content and GT-AG (donor/acceptor sites) structure was seen. However, a relatively high proportion of exon 1 along with intron 1 sequences were found to be possibly mis-annotated, and in most cases these were various ORFs in the 5′ intergenic regions joined with the correct acceptor site before exon 2. The success rate of the program in predicting exons, and the percentage of these predicted exons found to be within the correct size criterion (Fig. 3.4) indicate that, the overall success rate for identifying exons within the expected size ranges was: Exon 1: 48%, Exon 2: 88% and Exon 3: 54%.

Therefore, the overall ability of the programme to correctly identify an exon 1 and exon 3 was quite low, but was better at predicting a correct exon 2. The most commonly found causes for the misannotations of exon 1 and 3 also seemed to be different. For exon 1 many random ORFs were identified, whereas for exon 3 the donor site of the intron at the end of exon 2 was ignored, and the ORF of exon 2 continued for a few nucleotides into the second *yir* intron until a stop codon was eventually reached. The poor prediction for the exon 2/exon 3 splice site junction is an important finding as it is an ideal site for designing splice site containing primers. Based on this study, there is a high likelihood that primers designed based on the annotation data would be wrong. Overall, only around 54 % of the analysed *yir* genes seem to be predicted correctly by the algorithm, and emphasises that manual curation of *yir* genes based on experimental data as obtained in this chapter is needed.

Figure 3.4

Ability to detect exons and annotate them correctly

The three possible exon annotation groups named on the X-axis are shown. Two percentages are indicated; the first histograms (percent detected) show how many percent of the total 716 retrieved sequences were put into the three categories (exon1,2,3) based on manual identification. The percentages were: exon 1: 89%, exon 2: 100%, and exon 3: 63%. The second histogram (percent correct) shows how many percent within these three categories were estimated to be correctly annotated based on their size distribution in comparison to a number of manual annotations. The percentages were: exon 1: 54%, exon 2: 88%, and exon 3: 86%.


Percentage detected and annotated correctly
120
100
80
60
40
20
0
Exon1
Exon2
Exon3
Putative exon
🟦 Percent detected
🟥 Percent correct

Chapter IV

YIR phylogeny

4.1 Introduction

Prior to publication of the *P.yoelii* genome and its annotation (Carlton et al., 2002), 4 *yir* genes had been sequenced from *P.yoelii* gDNA (Janssen et al., 2000). These were deposited at EMBL under the accession numbers: AJ320478-AJ320481 and called *yir 1, yir 2, yir 3* and *yir 4* respectively (Janssen C 2001). When the *P.yoelii* 5x coverage sequencing was available, 838 *yir* genes were predicted. The average size of *yir* was 930 nt, with 90% of all *yir* being between 745 nt and 1115 nt (YIR protein average 310 aa, 90% between 248 aa and 372 aa). All *yir* open reading frames (ORF) were identified with Hidden Markov Models (HMM) and compared with the *yir* 1 to 4 sequences, as well as a published *P.berghei bir* gene and a *P.chabaudi cir* gene (Janssen et al., 2000 and Fischer 2003). From this comparison, the *yir* genes were classified as *yir* 1 to 4, *bir* or *cir* based on E value score and HMM analysis. This resulted in the *yir* sequences being assigned to each of the reference sequences as: *yir* 1: 158; *yir* 2: 92; *yir* 3: 262; *yir* 4: 194; *bir*: 119 and *cir*: 12. However, some of the original six reference sequences were published before the 5x coverage sequence data (Carlton et al., 2002), and it was considered likely that the reference sequences might not reflect the most divergent genes within the *yir* repertoire.

4.1.2 Objectives

Therefore, in order to be able to characterise the *yir* gene diversity and to provide a framework for the experimental studies, a phylogenetic analysis was performed on the sequences (4.2.1 to 4.2.5). The phylogenetic grouping was compared to the grouping suggested by TIGR (4.2.6). In addition, it was sought to establish how the *yir* genes in different phylogenetic groups varied in size, GC content and chromosomal localisation (4.2.7 to 4.2.12). A BLASTN based method was developed (4.2.13) to investigate whether a similar grouping of *bir* and *cir* genes existed in *P.berghei* and *P.chabaudi* (4.2.14 to 4.2.15). A set of very large *yir* genes could not be included in the initial phylogenetic analysis, and therefore the extent of this repertoire was investigated (4.2.16).

4.2 Results

4.2.1 Phylogenetic characterisation of YIR/*yir*

Translated sequences (YIR) and nucleotide sequences (*yir)* were analysed with various tree-building models: Neighbour-Joining (NJ), Minimum Evolution (ME) and Maximum Parsimony (MP). NJ and ME analysis were performed using Mega 2 and MP performed with PAUP (see materials and methods). Both the NJ and ME tree-building models are algorithmic distance methods and use the values in a distance matrix to compute branch order and lengths (Hall et al., 2004, 2nd edition). The NJ method is considered to be a simplified version of the ME method (Hall et al., 2004, 2nd edition), and does not investigate all possible tree topologies (Mega 2, description), whereas the ME method uses the least total branch lengths to generate tree topology. The MP method is a tree-searching character based method, which generates several different trees and decides the best based on a calculation of which tree involved the fewest changes (Hall et al., 2004, 2nd edition). Unlike the distance-based methods, which use the matrix file generated from an alignment, the MP method uses the alignment itself, and compares characters within each column of the alignment. It is beyond the scope of this thesis to discuss all the pros and cons of the different phylogenetic methods, but it should be emphasized that all trees are only evolutionary models, which relies on certain assumptions that might or might not be a reflection of actual evolutionary events. However, the aim in this thesis was to generate a framework, whereto experimental results could be related, and for this the NJ, ME and MP methods seemed reasonable choices.

The phylogenetic analysis was performed as described (Fig. 4.1, and materials and methods). Briefly (Fig. 4.1), a comparison between NJ and ME protein grouping was performed, and the NJ protein grouping were compared with nucleotide groupings. The nucleotide groupings were then compared between NJ and ME and NJ and MP models, and a consensus grouping was created, where sequences found to be consistently different between the three models were removed from the grouping. Bootstrap consensus NJ trees were constructed to evaluate the validity of the phylogeny and to determine whether the groups could be merged into supergroups.

4.2.2 Phylogenetic analysis of translated sequences

From an alignment of all 838 *yir* genes, all exon 3 sequences were removed due to misannotations (discussed in Chapter III). The genes were then translated using the standard genetic code. All sequences containing multiple stop codons were removed. An initial attempt to construct an NJ tree failed because some translated sequences were too divergent (NJ analysis in Mega 2 requires in the alignment at least one column with 100% similar residues). These divergent sequences were also removed and this resulted in 718 YIR proteins for phylogenetic analysis, which constitute 86% of the YIR repertoire.

These sequences were aligned and used for phylogenetic analysis. Initially (Fig. 4.2 a), nine different protein groups (Pg 1-9) could be identified using the NJ method. The grouping was highly reproducible when using translated sequences and a ME tree building method (Fig. 4.2 b). Also, using nucleotide sequences and a NJ tree building method (Fig. 4.2 c) reproduced the grouping, although especially one group (Pg 5) appeared to be more scattered on this tree. Overall, a similar grouping could

Figure 4. 1

Flowchart for the phylogenetic analysis

Both the translated amino acid (aa) and nucleotide sequences (nt) from 718 annotated *yir* exon 1 and 2 were used to generate two multiple alignments. The multiple aa alignment was used to generate two trees using the Neighbour Joining (NJ) and Minimum Evolution (ME) models. Visually identifiable branch points on the aa NJ tree were used to identify groups, and the aa ME tree was then evaluated for a similar clustering of the same groups. The aa NJ groups were then compared to the nt NJ tree, which were used as the reference point for further comparisons. In comparison a), visually identifiable groups on an NJ tree (distance and topology) were evaluated on the ME nt tree also (distance and topology). In comparison b), the groups on the nt NJ tree were evaluated on a nt Maximum Parsimony (MP) tree. The NJ group model was evaluated on the basis of comparison a) and b). From this, five supposed supergroups defined from the DNA NJ tree are evaluated by clustering on the ME and MP trees. Finally the supergroups were evaluated by consensus tree analysis, which uses the bootstrap values to test how reproducible the inferred tree grouping was.

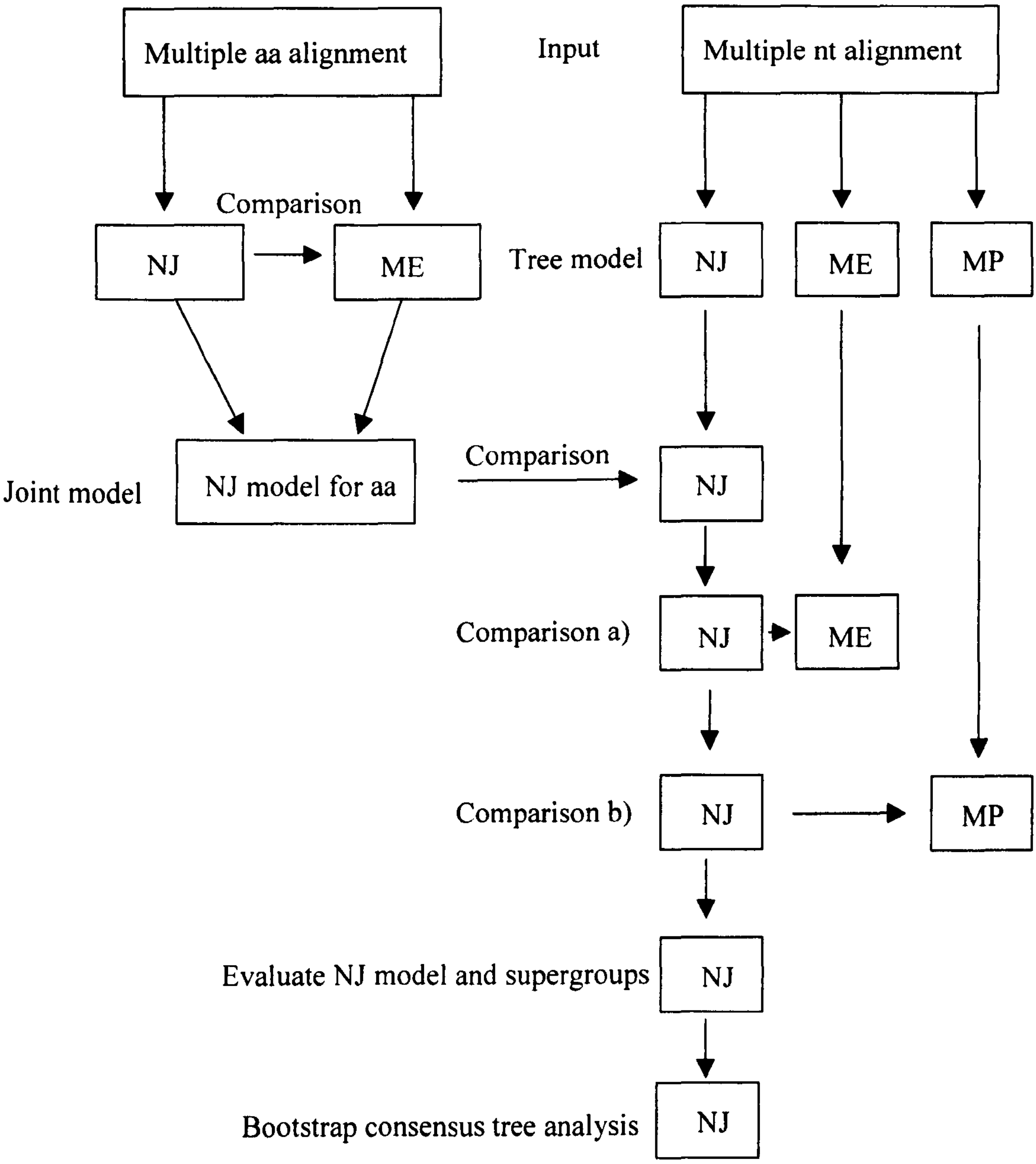

Multiple aa alignment
Input
Multiple nt alignment
Comparison
NJ
ME
Tree model
NJ
ME
MP
Joint model
NJ model for aa
Comparison
NJ
Comparison a)
NJ
ME
Comparison b)
NJ
MP
Evaluate NJ model and supergroups
NJ
Bootstrap consensus tree analysis
NJ

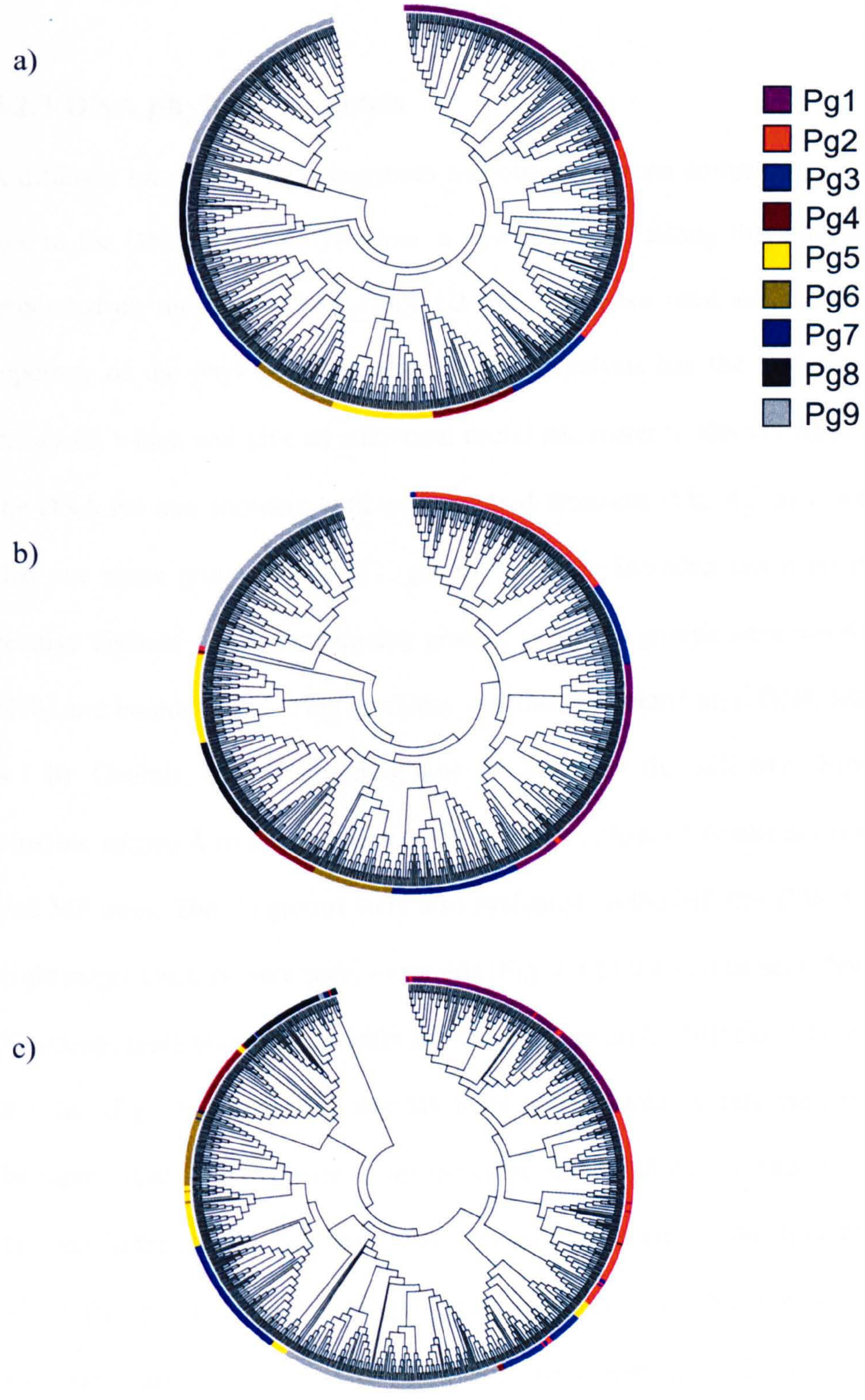
Figure 4.2

Groups on protein NJ/ME and DNA NJ trees

a) Visual inspection of obvious branching points on the protein NJ tree suggested the existence of nine protein groups (Pg1 to Pg9), which were colour coded accordingly.

b) The locations of the nine groups (suggested in a) were investigated on a protein ME tree.

c) The locations of the nine groups (suggested in a) were investigated on a nucleotide NJ tree.


a)
Pg1
Pg2
Pg3
Pg4
Pg5
Pg6
Pg7
Pg8
Pg9
b)
c)

be observed regardless of whether amino acid or nucleotide sequences, or if NJ or ME tree building methods were used.

4.2.3 DNA phylogeny models

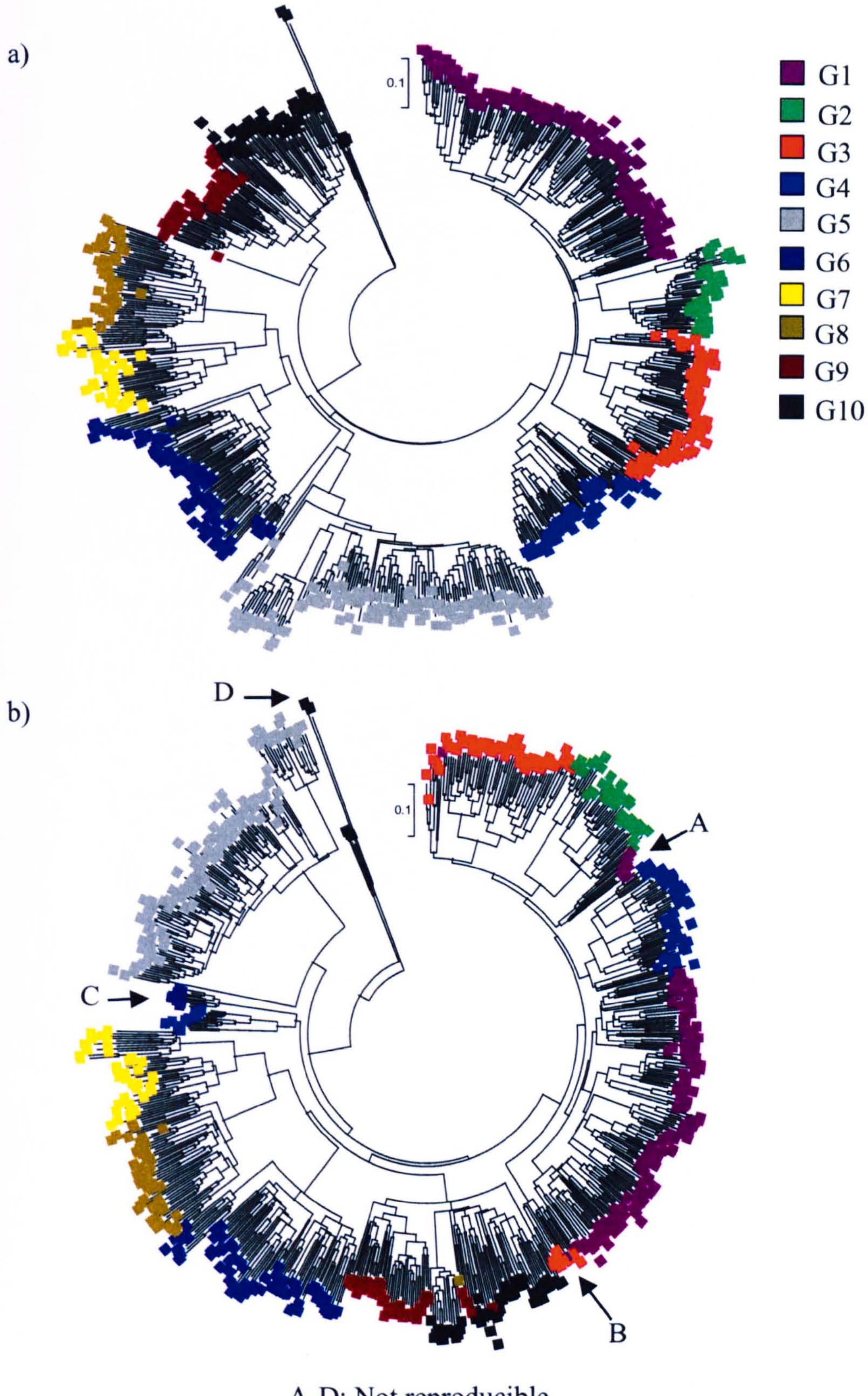
A different location of Pg5 sequences was observed when comparing the protein NJ tree to the DNA NJ tree. Therefore, a new grouping, taking this into account was generated on the basis of the DNA NJ tree. The trees used so far only show the topology of the phylogeny, however Mega 2 analysis has the ability to show the distances, which will give an additional useful parameter in identifying groups. From the DNA NJ tree showing both topology and distances (Fig. 4.3 a) it was apparent that one more group, called G2 (green), could be included based on its different relative distance to the surrounding groups. In all, 10 groups were identified (G1 to G10) and colour coded. This grouping was then evaluated on a DNA ME tree (Fig. 4.3 b). Overall, the NJ grouping was the same on the ME tree. However, four clusters, named A to D consisting of 45 sequences changed positions between the NJ and ME trees. The NJ groups were also evaluated on the MP tree (Fig. 4.4 a), where eight major clusters were seen. From this (Fig. 4.4 b) it could be seen that overall the NJ groups were maintained within specific clusters on the MP tree. The evaluation of the ten NJ groups on the ME and MP trees revealed that overall, the grouping could be reproduced. The majority of the differences that did exist could be explained by changes in the relative positions of groups within five major branch points (Fig. 4.5). These five branch points were thought to represent a possible supergrouping of the individual groups. The individual groups were merged into the five suggested supergroups (SG1 to SG5) (group 5 and 6 were retained, but renamed as SG2 and SG3 respectively).

Figure 4.3

DNA topology-distance trees

a) DNA NJ tree. As in figure 4.2 c, but also showing distances between the groups by positioning them in an outwards manner, reflecting their relative distances. A total of ten groups were identified. All these both shared a branch (topology), and were located in the same relative outwards position (distance). These are colour coded as shown in the key.

b) DNA ME tree. The ten DNA groups were colour coded accordingly to above (a). Four clusters of sequences (45), indicated by bold arrows named A to D, were not reproducible between the two tests.



A-D: Not reproducible

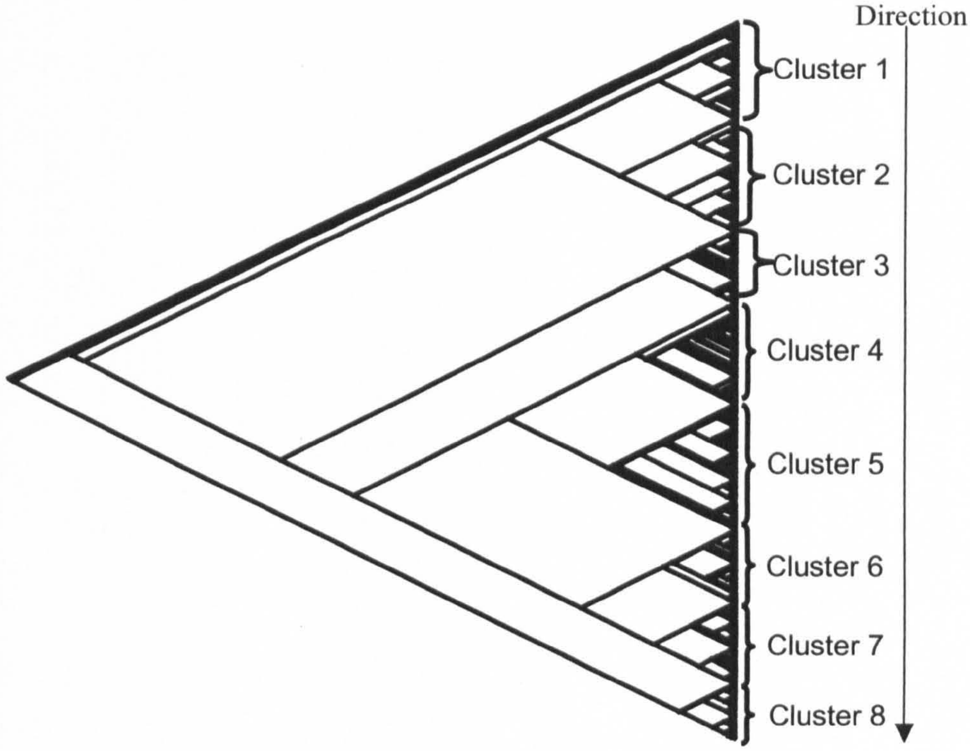
Figure 4.4

Maximum Parsimony tree

a) Maximum Parsimony tree generated in PAUP and viewed in TreeView (see materials and methods). In TreeView, groups could not be colour coded, and therefore the location of eight visually identifiable clusters are indicated, in addition to an arrow indicating the direction on the tree any following analysis refer to.

b) Evaluation of the NJ groups on the MP tree. In the graph, the x-axis corresponds to the direction of the arrow shown above (a). On the y-axis is indicated which of the ten NJ groups were located at a particular location on the MP tree. The following NJ groups were observed within the eight clusters: Cluster 1: NJ group 1 to 3, Cluster 2: NJ group 1 and 4, Cluster 3: NJ group 9 and 10, Cluster 4: NJ group 5, Cluster 5: NJ group 6, Cluster 6: NJ group 8, Cluster 7: NJ group 7, Cluster 8: NJ group 3.

a)


Direction
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8

b)

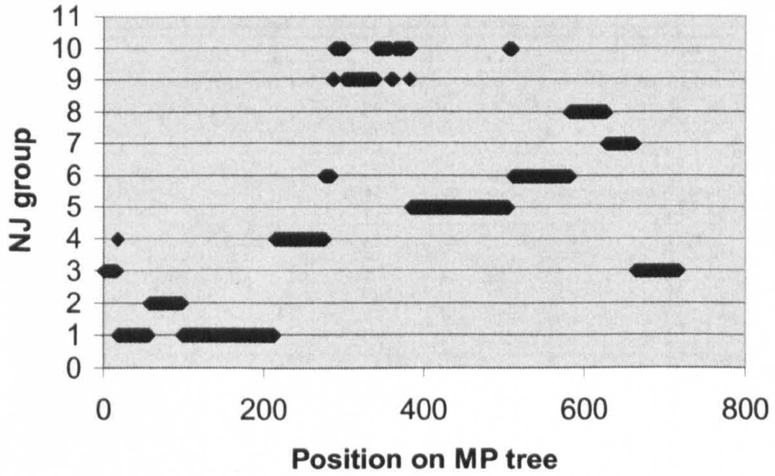

11
10
9
8
7
6
5
4
3
2
1
0
NJ group
0
200
400
600
800
Position on MP tree

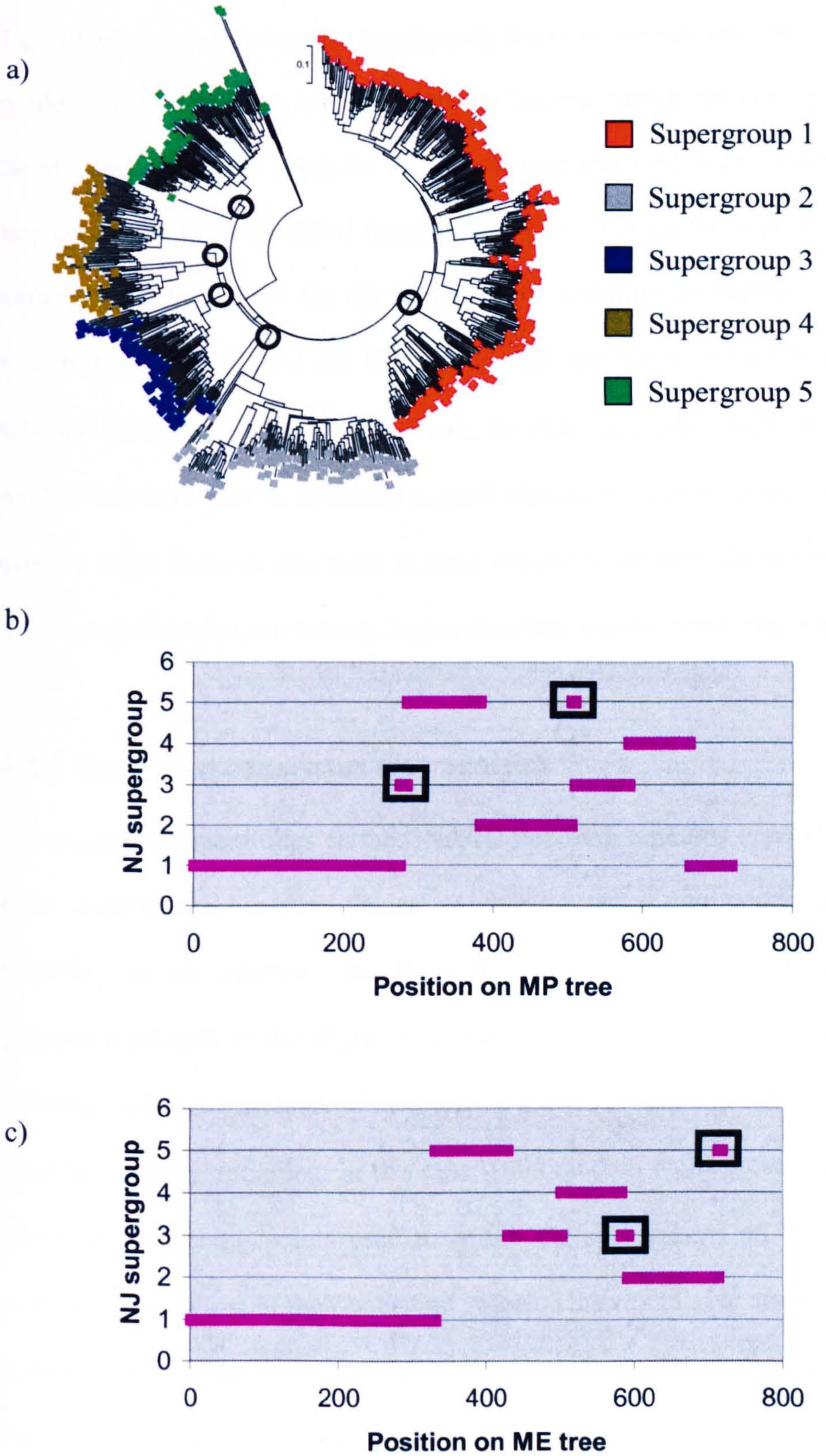
Figure 4.5

Comparison of suggested supergrouping

a) The proposed supergroups (SG1 to SG5) consistent with most of the minor rearrangements are seen on the NJ tree and colour coded as indicated in the legend.

b) Evaluation of the supergroups (SG1 to SG5) on the MP tree. The x-axis shows the position on the MP tree and the y-axis shows the five supergroups. Some sequences were located outside of the supergroups, and these are indicated by black boxes.

c) Evaluation of the supergroups (SG1 to SG5) on the ME tree. The x-axis shows the position on the ME tree and the y-axis shows the five supergroups. Some sequences were located outside of the supergroups, and these are indicated by black boxes. These sequences were the same identified in 4.5 b.


a)
0.1
🟥 Supergroup 1
▨ Supergroup 2
■ Supergroup 3
■ Supergroup 4
🟩 Supergroup 5
b)
NJ supergroup
0
1
2
3
4
5
6
0
200
400
600
800
Position on MP tree
c)
NJ supergroup
0
1
2
3
4
5
6
0
200
400
600
800
Position on ME tree

It was then investigated how well the supergrouping would compare with clusters on the ME and MP trees. The NJ supergroups remained clustered on both the MP (Fig. 4.5 b) and ME (Fig. 4.5 c) trees. SG1 was split into two locations on the MP tree (Fig. 4.5 b), but as this was not consistently found on the ME tree, SG1 was retained. In addition, 14 sequences were consistently located outside the supergroups on both the MP and ME trees. These 14 sequences were removed from further analysis, as they could not be categorized consistently. The order of the supergroups differed somewhat on the 3 trees; On the NJ tree (used to define the supergroups) the order was: SG1, SG2, SG3, SG4 and SG5. On the ME tree the order was: SG1, SG5, SG3, SG4 and SG2, and finally on the MP tree, the order was: SG1, SG5, SG2, SG3, SG4. As SG1 and SG5 were consistently located adjacent to each other on two out of three trees, it might indicate that there is some similarity between the sequences in these two supergroups that is relatively higher than between the remaining supergroups.

4.2.4 Bootstrap consensus tree analysis

The suggested supergroups on the DNA NJ tree were tested by generating consensus trees based on the bootstrap values. A bootstrap test is commonly used to test the reliability of an inferred tree. Basically, this involves random sampling of a proportion of each of the aligned sequences and then regenerating the tree, and the bootstrap value is a measure of how often a particular branch could be reconstructed from this random sampling. In this case, 1000 random replications were performed. The outer branching points used to define the supergroups on the NJ tree were evaluated by looking at their bootstrap values. This would give an indication of how reliable the supergrouping was. Bootstrap consensus NJ trees were generated at bootstrap values of 51% and 86% (Fig. 4.6). At these bootstrap values, SG1 collapsed at bootstrap values above 51% (Fig. 4.6 a and b), whereas the remainder of

Figure 4.6

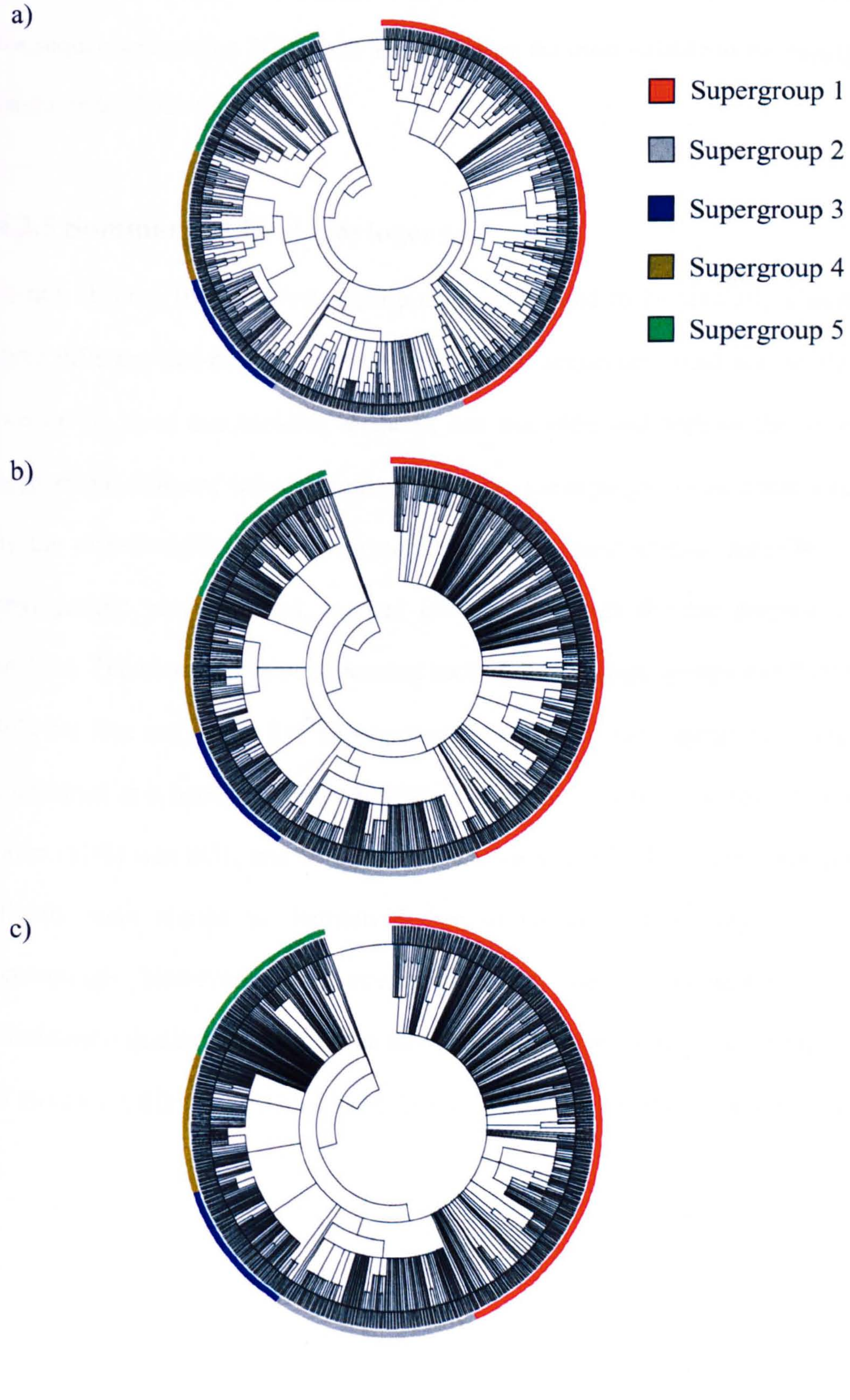
Bootstrap consensus trees

For the DNA NJ tree with the five supergroups indicated, ranges of different bootstrap values were set and the resulting trees are shown. If the bootstrap value for a branch is below the set value, the result is seen as polytomies.

a) NJ tree at 51% bootstrap value. At this value, the outer branching points defining SG1 to SG5 is retained.

b) NJ tree at 86% bootstrap value. At this value, SG1 had developed into polytomies, whereas SG2 to SG5 are all retained.

c) NJ tree at 95% bootstrap value. At this value only SG2, SG3 and SG4 still retain their outer branching points.


a)
b)
c)
🟥 Supergroup 1
⬜ Supergroup 2
🟦 Supergroup 3
🟫 Supergroup 4
🟩 Supergroup 5

the supergroups were kept at bootstrap values up to 86% (Fig. 4.6 b). At 95% bootstrap values (Fig. 4.6 c), only SG2, SG3 and SG4 kept their outer branching points. This meant that the sequences in SG1 were less reliable than the remainder of the sequences and that SG2 , SG3 and SG4 were the most reliable of the supergroups based on the bootstrap values.

4.2.5 Summary of DNA phylogeny

In this chapter, five different supergroups were found to be similarly clustered in three different tree-building models. A set of 14 sequences could not be placed in two of the three tree building methods, and two (ME and MP) of the three trees suggested a different order of similarity between the supergroups than that suggested by the reference NJ tree. The identification of the same groups, regardless of the phylogenetic program used, seemed consistent enough for the purpose of this analysis. The bootstrap value separating each of the five supergroups was found to be 86% for four out of the five supergroups, and some of the supergroups were even maintained at a bootstrap value of 95%. The supergroup with the lowest bootstrap value (51%) was SG1, and it is possible that *yir* genes in this supergroup are only slightly more similar to themselves than to *yir* genes from any of the other supergroups. However, SG1's internal similarity was still enough to give it a consistently similar clustering in the three tree building methods, and for the purpose of this thesis, SG1 was deemed to be different enough to maintain it as a supergroup.

In short, the following was established in this chapter:

- 86% of the *yir* repertoire can be divided into 10 groups, each containing the following number of genes: G1: 153, G2: 141, G3: 73, G4: 64, G5: 121, G6: 81, G7: 36, G8: 45, G9: 49, G10: 55.
- These 10 groups can be reduced to five supergroups by merging G1 to G4 into SG1 (331 genes), G5 kept as SG2 (121 genes), G6 kept as SG3 (81 genes), G7 and G8 merged into SG4 (81 genes) and G9 and G10 merged into SG5 (104 genes).
- Bootstrap values for the five supergroups ranges from 51% (SG1), 86% (SG5), 95% (SG2, SG3 and SG4).

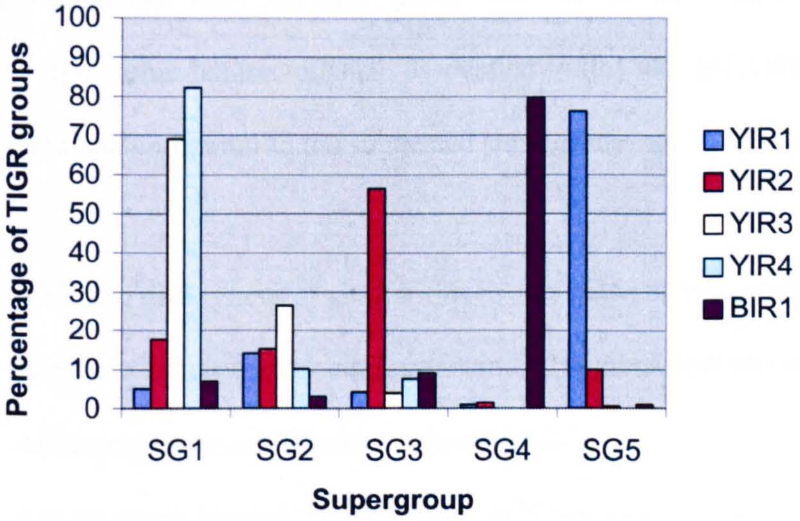
4.2.6 Comparison of supergroups to TIGR annotation

A comparison between the Supergroups (SG1-SG5) and the TIGR grouping (*yir* 1 to 4 and *bir*) was made (Fig. 4.7). The 12 *cir* like sequences were omitted from this analysis due to their low number, but most of them (67%) belonged to SG2. To summarize: 80% of the *bir* like genes were in SG4, 76% of the *yir1* genes were in SG5, 82% of the *yir4* and 69% of the *yir3* genes were in SG1 and 56% of the *yir2* genes were in SG3. This showed that some of the TIGR groups matched better with the supergroups than others; SG1, SG4 and SG5 contained at least 69% of one of the TIGR annotation groups. However even for these, between 18 and 31 % of the TIGR annotation groups were located outside of the major supergroup, and located within SG2 and SG3 for the most part. It seemed as if SG3 had been partially missed by the TIGR annotation groups. SG2 had been completely missed, as it contains a random low percentage mix of various TIGR annotation groups. Since both SG2 and SG3

Figure 4.7

Comparison of *yir* supergroups to TIGR grouping

The percentages of each of the TIGR annotation groups, which were present within each of the five supergroups, were calculated. SG1, SG4 and SG5 contained approximately 80% of the TIGR annotation groups *yir4*, *bir1* and *yir1* respectively. Two small TIGR annotation groups, consisting of 12 sequences called *cir1* and *cir2* were omitted from this analysis, but 8 out of the 12 sequences were found within supergroup 2.


Percentage of TIGR groups
100
90
80
70
60
50
40
30
20
10
0
SG1
SG2
SG3
SG4
SG5
Supergroup
YIR1
YIR2
YIR3
YIR4
BIR1

had 95% bootstrap values in the previous analysis, these supergroups contained sequences that were among the most distinct within the *yir* repertoire. The most likely reason for the missing TIGR annotation of sequences within these supergroups was probably due to the absence of representatives of these supergroups in the reference sequences. Therefore, this current analysis represents an improvement in classification of the *yir* repertoire.

4.2.7 What characterises the groups and supergroups?

Having established that the *yir* genes can be divided distinct groups (and supergroups), some characteristics associated with the individual groups were investigated and compared to the suggested supergrouping:

1) Are there differences or similarities in size between the groups?
2) Are there differences or similarities in GC content between the groups?
3) Are some *yir* genes located centromerically?
4) Are *yir* genes located "more or less" subtelomerically?

4.2.8 Are there differences in sizes between the groups?

The ten groups identified by the phylogenetic analysis could contain *yir* genes of different sizes. To analyse this, the sizes of all the 718 *yir* genes used in the phylogenetic analysis were calculated. The size of each individual gene within each of the groups was plotted (Fig. 4.8 a), and from this it could be seen that some of the groups, especially group 3, contained genes with markedly different sizes. The average size for all genes in each of the groups (Fig. 4.8 b) showed that the groups followed two major size ranges, G1, G2, G4, G5 and G6 between 775-825 (nt), and G3, G7, G8, G9 and G10 between 875-925 (nt).

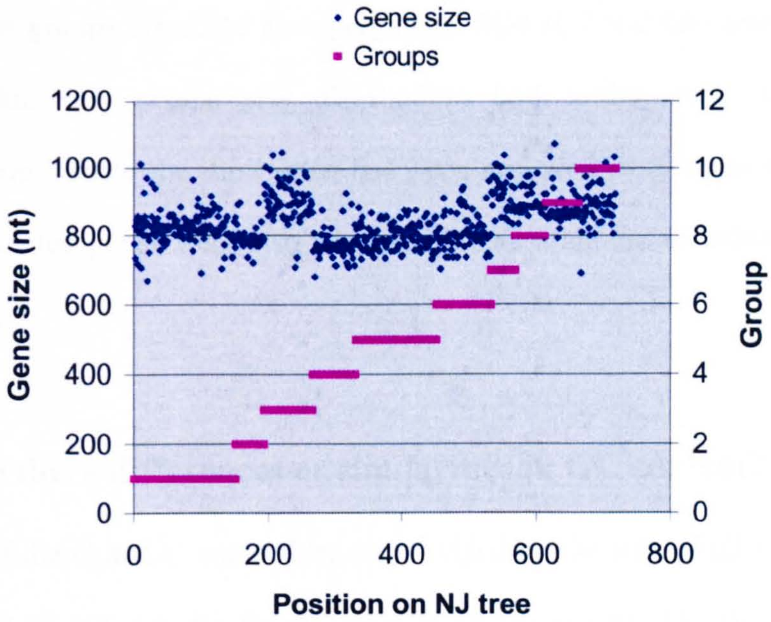
Figure 4.8

Size distributions of genes in groups

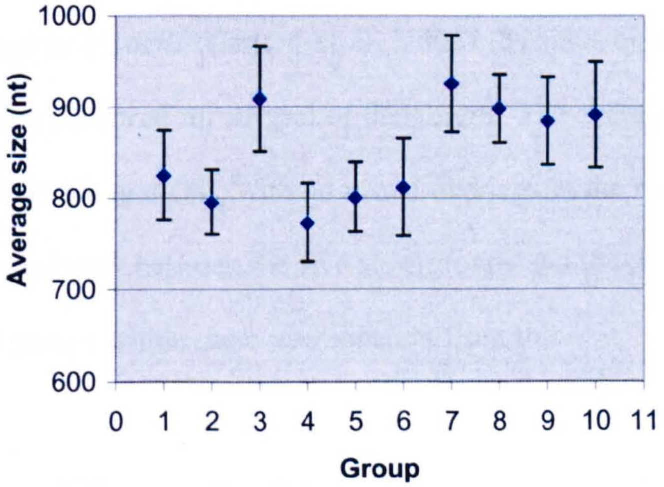
a) Size distribution among groups. On the X-axis is shown the position on the DNA NJ tree with the ten groups indicated as solid bars. On the left Y-axis is indicated the size in nt, and on the right Y-axis the group number.

b) The average sizes and standard deviations for the 10 groups (on the X axis) and with the size in nt on the Y-axis. The median gene size in the individual groups were (all in nt): G1: 825, G2: 795, G3: 908, G4: 773, G5: 801, G6: 812, G7: 925, G8: 898, G9: 884, G10: 891. The median values of the groups were compared with Dunn's Multiple Comparison test in Prism. The following groups were found to be statistically different (at either P<0.01: ** or P<0.001: **): **G1**/G2,G3,G4,G5,G7,G8,G9,G10. **G2**/G3,G7,G8,G9,G10. **G3**/G4,G5,G6; **G4**/G6,G7,G8,G9,G10. **G5**/G6,G7,G8,G9,G10. **G6**/G7,G8,G9,G10.

a)


• Gene size
▪ Groups
Gene size (nt)
1200
1000
800
600
400
200
0
Group
12
10
8
6
4
2
0
0
200
400
600
800
Position on NJ tree

b)


Average size (nt)
1000
900
800
700
600
0
1
2
3
4
5
6
7
8
9
10
11
Group

nt. Comparing these size distributions within the five supergroups (SG1 to SG5) showed that SG1 (consisting of G1 to G4) contained groups, which were highly variable in sizes. SG2 (G5) and SG3 (G6) could not be split up into groups as they were in fact groups classified as supergroups. SG4 (G7 and G8) and SG5 (G9 and G10) contain very similar size distributions both within each supergroup and between them. Since the third exon had been removed from these sequences, full length annotated genes were also investigated and a similar distribution was found (not shown).

4.2.9 Are there differences or similarities in GC content?

The sum of the G and C nucleotides was divided by the sum of all four nucleotides (A,G,C and T) individually for all genes in all ten groups. The GC percentage for each individual gene was plotted against the group it belonged to (4.9 a), and for all genes in each group, 99% confidence intervals were calculated (4.9 b) In total the average GC content of all analysed *yir* genes was 25.7% compared to 24.8% for all annotated genes in *P.yoelii* (Carlton et al., 2002). Distinct distributions in average GC content were observed for several of the groups. The average GC content varied by up to 2 % (G1/G4 and G8), with no visual overlaps in the repertoire (Fig. 4.9 b). However, no similarity between the five supergroups and the average GC content of the individual groups within them was apparent from this.

The only trend, which could be observed from this, was that the two groups with the lowest average GC content (G1 and G4) were in SG1, whereas the two groups with the highest GC content (G8 and G9) were in SG4 and SG5 respectively.

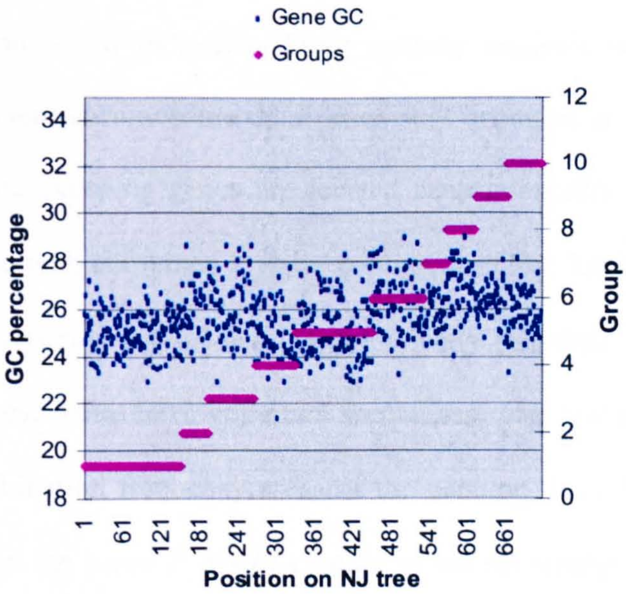
Figure 4.9

GC percentages of genes in groups

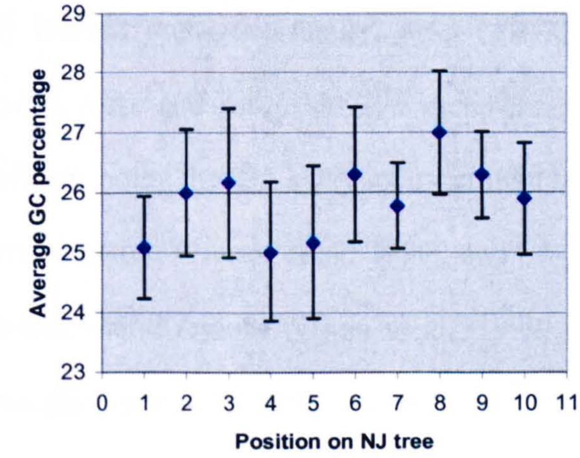
a) Distribution of GC percentages for genes in the ten groups. On the X-axis is shown the position on the DNA NJ tree with the ten groups indicated as solid bars. On the left Y-axis is indicated the GC percentage calculated for each individual gene, and on the right Y-axis the group number.

b) Median GC percentages and standard deviations for the 10 groups (on the X axis) and with the GC percentage on the Y-axis. The median GC percentages of the individual groups were: G1: 25.19% G2: 26.08 %, G3: 26.46% and G4: 25.16%, G5: 25.20% and G6: 26.59%, G7: 25.66%, G8: 27.03%, G9: 26.32% and G10: 26.01%. The median values of the groups were compared with Dunn`s Multiple Comparison test in Prism. The following groups were found to be statistically different (at either P<0.01: ** or P<0.001: **): **G1**/G2,G3,G6, G8,G9,G10. **G2**/G8. **G3**/G4, G5. **G4**/G6,G8, G9, G10. **G5**/G6,G8,G9. **G7**/G8. **G8**/G10.

a)


Gene GC
Groups
34
32
30
28
26
24
22
20
18
GC percentage
12
10
8
6
4
2
0
Group
1
61
121
181
241
301
361
421
481
541
601
661
Position on NJ tree

b)


29
28
27
26
25
24
23
Average GC percentage
0
1
2
3
4
5
6
7
8
9
10
11
Position on NJ tree

4.2.10 Are some *yir* genes located centromerically?

In order to see if there was any physical compartmentalization of the genes in the different supergroups, it was first investigated if any yir genes shared contigs with housekeeping genes. Due to the lack of chromosome data for *P.yoelii*, this question can only be answered indirectly. From synteny analysis between *P.yoelii* and *P falciparum*, where chromosome data do exist (Carlton et al., 2002 and 2005), it is known that housekeeping genes are located centromerically and that their order is highly syntenic with combined *P.yoelii* contigs. It is also known that 35% of the *var* genes in *P.falciparum* are located centromerically (Gardner et al., 2002), and it is thought that these genes have important specialised roles and are therefore not placed in high recombination frequency areas of the genome. It is therefore likely that *yir* genes found on the same *P. yoelii* contigs as housekeeping genes are also located centromerically. By analysing all the *yir* containing contigs for the presence of non-*yir* assigned genes it was found that *yir* genes share contigs with 660 hypothetical genes and 5 genes with an assigned function. The genes with an assigned function were a gene assigned as "93Kda protein" (PY01007, PY01334 and PY04884), *py235* (PY01185) and a lysophospholipase (PY01001). The latter is a member of another subtelomerically located multigene family, *pst-a* (TIGR01607 HMM ID), which, together with *pst-b*, *pst-c* and *pst-d*, has 275 members (Carlton et al., 2002). The *py235* gene family encoding the 235 kDa rhoptry protein has been reported to exhibit phenotypic clonal variation (Preiser et al., 1999), and has previously been shown to be located in subtelomeric regions (Owen et al., 1999). No information could be found for the assigned gene "93 KDa protein", but its subtelomeric location was indirectly shown as one of the genes (PY01007) was found to be located on a 20 KB contig (MALPY00270) containing both a *yir* gene (PY00997) and the lysophospholipase gene (PY01001). Since there were 385 more hypothetical genes than can be accounted for by the 275 members of the *pst-a/b/c/d* families (most of which are annotated as hypotheticals) it opens up the potential for the presence of other large, yet uncharacterised, multigene families in *P. yoelii.*

Of importance here is that the identified genes with an assigned function, could all be ascribed as being located subtelomerically. It was found that 58 % (4498 out of 7752) of all annotated genes in the genome were assigned as hypotheticals and most of the remaining non-*yir* genes were housekeeping genes thought to be located centromerically. This contrasts with the findings for the *yir* containing contigs, where 99.2 % (655 out of 660) of the genes were hypotheticals. This suggests that the *yir* genes analysed here do not share contigs with genes expected to be located in the centromeric regions.

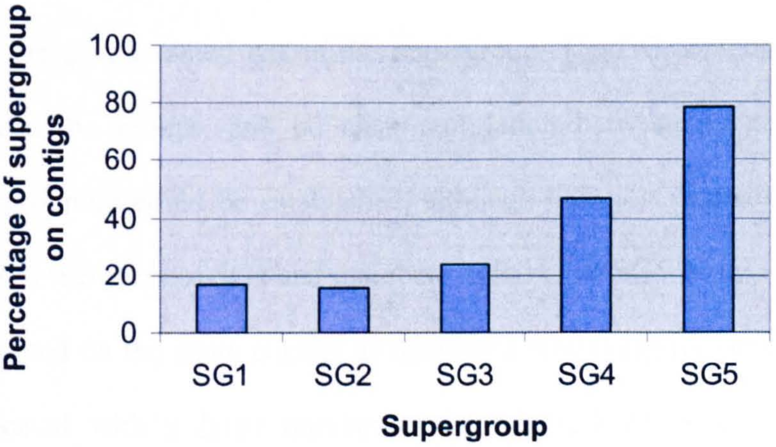
4.2.11 Are *yir* genes located "more or less" subtelomerically?

The designation "subtelomeric region" refers to regions in the immediate vicinity of the telomeric repeats at the end of chromosomes. However, the *yir* gene family has 838 members and even a low estimate of the size needed for a *yir* gene and its surrounding UTRs of 1.5 Kbp would yield 1257 Kbp of genomic sequence solely occupied by *yir* genes. Even this is a gross underestimate, as it was documented above that at least 665 hypothetical genes shared contigs with the *yir* genes. However this analysis also clearly showed that the *yir* genes are located in regions not shared by housekeeping genes, known by synteny analysis to be located centromerically. If the subtelomeric regions are distributed equally among the 28 chromosome ends, each subtelomeric region would be at least 44 Kbp, which could be divided into discrete chromosomal territories. Based on this it is very likely to use the term "subtelomeric" to describe regions, where multigene families resides to distinguish them from regions where housekeeping genes resides. TIGR had assigned some contigs as telomeric and subtelomeric contigs (Carlton et al., 2002), identified by the presence of telomeric repeats (AACCCTG –found on 71 contigs, reduced to 28 after gap closure) and other repeats identified through MUMmer 2 and Tandem Repeat finder. Especially a 15 bp repeat with 45 copies and a 31 bp repeat with up to 10 copies (not shown in their analysis) were found on 271 individual contigs, although no Rep20 elements –found in *P.falciparum* were found (Carlton et al., 2002). These telomeric and subtelomeric contigs were used to analyse the distribution of the *yir* groups on these, as this would indicate a "more or less" subtelomeric localisation of the groups. The contigs annotated by TIGR were retrieved from http://www.tigr.org, and these contigs were then analysed for the location of *yir* genes. The result of this analysis (Fig. 4.10) clearly indicates that a striking distribution of supergroups on these subtelomeric contigs exists. The percentages of each supergroup located on these contigs increases almost linearly from SG1 to SG5, and this also reflects the original clockwise supergroup order on the DNA NJ tree. It should be noted that G3 (a component groups of SG1) accounted for 40% of the subtelomerically located SG1 genes although it only comprised 22% of SG1's repertoire. Since G3 genes were also larger than the remaining SG1 genes (see Fig. 4.8 a), there is a correlation between the size distribution and localization, as the larger SG4 and SG5 genes were the most subtelomerically located of the supergroups. However, this was not consistently supported by the phylogeny, as G3 genes resembled genes within SG1 more than genes in SG4 or SG5.

Figure 4.10

Subtelomeric localisation

Annotated subtelomeric and telomeric contigs were obtained from TIGR (see materials and methods). The contig number was used to identify which contig, *yir* genes were located on, and the supergroup identity of each *yir* was established. The graph depicts this as the percentage of *yir* genes from each of the supergroups (out of that supergroups total number of *yir* genes) that were located on the annotated subtelomeric/telomeric contigs.


Percentage of supergroup on contigs
100
80
60
40
20
0
SG1
SG2
SG3
SG4
SG5
Supergroup

4.2.12 Summary of size, GC content and localization

Genes from G7 to G10 were generally larger than genes from the other group, except G3, which contained the largest genes. Since G7 to G10 form SG4 and SG5, size difference is therefore one defining criterion that distinguishes these two supergroups from SG1 to SG3. However, that size alone was not what has defined the supergroups in the phylogenetic analysis can be seen from G3, which is a component of SG1, and yet contains genes, which are different in size compared to genes in the other groups contained within this supergroup. The GC content was highly variable between the groups, and no clear correlation between average GC content and supergrouping could be established, although the most extreme differences existed between members of SG1 and members of SG4 and SG5. No *yir* genes were found to be located on the same contigs as annotated housekeeping genes, instead they were co-located with a large number of hypothetical genes as well as other genes characterised as subtelomerically located.

The proportion of genes in each supergroup located on subtelomeric contigs (SG1: 17%, SG2: 16%, SG3: 23%, SG4: 47% and SG5: 78%) correlates strongly with the clockwise location of the supergroups on the NJ-DNA tree. In short, the following was found through this series of analysis of what characterises the groups and supergroups:

- The individual groups consisted of genes with distinct sizes. Of the supergroups that consisted of more than one individual group, SG4 and SG5 contained genes with very similar sizes, whereas SG1 contained genes with highly divergent sizes

- The GC content of genes within the groups did not clearly relate to the supergrouping, however the most extreme differences in GC content was observed for genes within SG1 and SG4+SG5 with the latter having a higher GC content.
- No *yir* genes co-localized with annotated housekeeping genes on the contigs, but instead co-localized with numerous hypothetical genes and a few subtelomerically located genes.
- There was a clear correlation between the proportion of genes located on subtelomeric and telomeric contigs and the five supergroups. This distribution followed the NJ-tree supergrouping in a clockwise manner so SG1 had the lowest proportion of genes on the subtelomeric and telomeric contigs, whereas SG5 had the highest.

4.2.13 Comparative genomic analysis

The two other rodent malaria species *P.berghei* and *P.chabaudi* contain the *yir* homologoues *bir* and *cir* respectively (Janssen et al., 2000 and Fischer et al., 2003). These had, together with the *yir* of *P.yoelii*, been grouped into the TIGR family, TIGR001590 (HMM ID). With 245 annotated *bir* and 130 *cir* genes (TIGR annotation data) the repertoires of *P.berghei* and *P.chabaudi* was significantly smaller than the 838 *yir* genes in *P.yoelii* (Carlton et al., 2002). These three rodent malaria species are all closely related, with *P.berghei* and *P.yoelii* being the most closely related based on cytochrome c phylogenies (Perkins et al., 2002). Having established that the *yir* genes can be put into five supergroups, it would be interesting to see if this supergrouping was evolutionary conserved in the two other rodent malaria species. If it was, the *yir* genes could be regarded as a likely expansion of an already established set genes shared between the rodent malarias. If this supergrouping was unique for *P.yoelii*, not only enlargement, but also diversification of the repertoire had occurred in this species from the last common ancestor. Depending on the exact evolutionary relationships between the rodent malarias, a restriction of the *bir* and *cir* repertoires could also have occurred.

To do this, a BLASTN analysis was performed on the genome of the two species, thus avoiding the problems of incomplete gene annotations. To perform a BLASTN analysis, a highly representative reference sequence was created from individual alignments of genes in the five supergroups. From these alignments, five 20 % consensus sequences were generated. A 20 % consensus sequence was chosen, as higher stringencies might not detect possible shared semi-conserved regions within each supergroup. As 14 *yir* genes could not be consistently located on the phylogenetic trees, these were removed from this and all subsequent analyses. The five consensus sequences were aligned (See S4.1) and the overall level of conservation between the five sequences were analysed. From a conservation plot of the consensus sequences, divided into 19 regions (R1 to R19, Fig. 4.11) it can be seen that there is a lack of conservation between R15 and R17. This corresponds to longer inserted sequences in the SG4 and SG5 consensus sequences. These regions are located in the vicinity of the transmembrane encoding end of the *yir* exon 2. Therefore, there are differences between these consensus sequences that would make them suitable to function as reference sequences. The reliability of the reference sequences in detecting only genes from their own supergroup with a high significance level was tested through BLASTN analysis. This was done to determine whether sets of superhomologues genes could be identified through a shared low E value when compared with a reference sequence generated from each of the supergroups.

Figure 4.11

Conservation plot for SG1 to SG5 consensus sequences

The five aligned 20% consensus sequences (See S4.1) were divided into 19 regions, called R1-R19, consisting of 50 nucleotides each, except for the last region, which consisted of only 41 nucleotides. For each region, it was calculated how many nucleotides were 100% conserved. The 19 regions are shown on the x-axis, and the percentage conservation within each region is indicated on the y-axis. The regions Rl to R19 correspond to the arbitrary 5′ to 3′ directions on a *yir* gene.

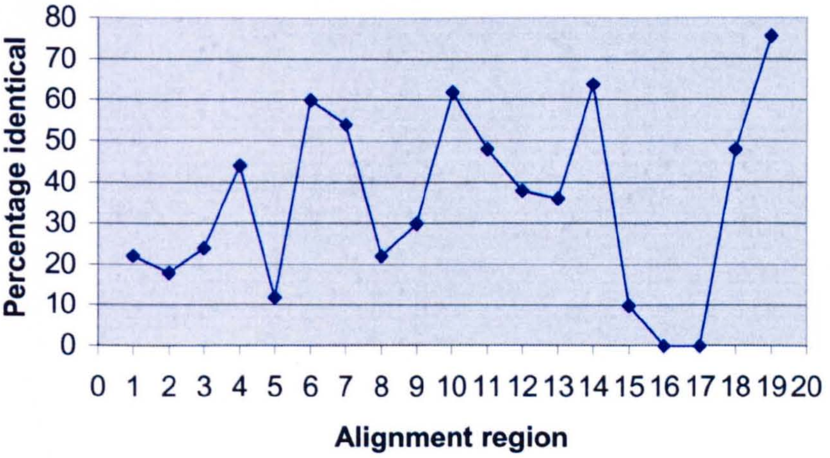
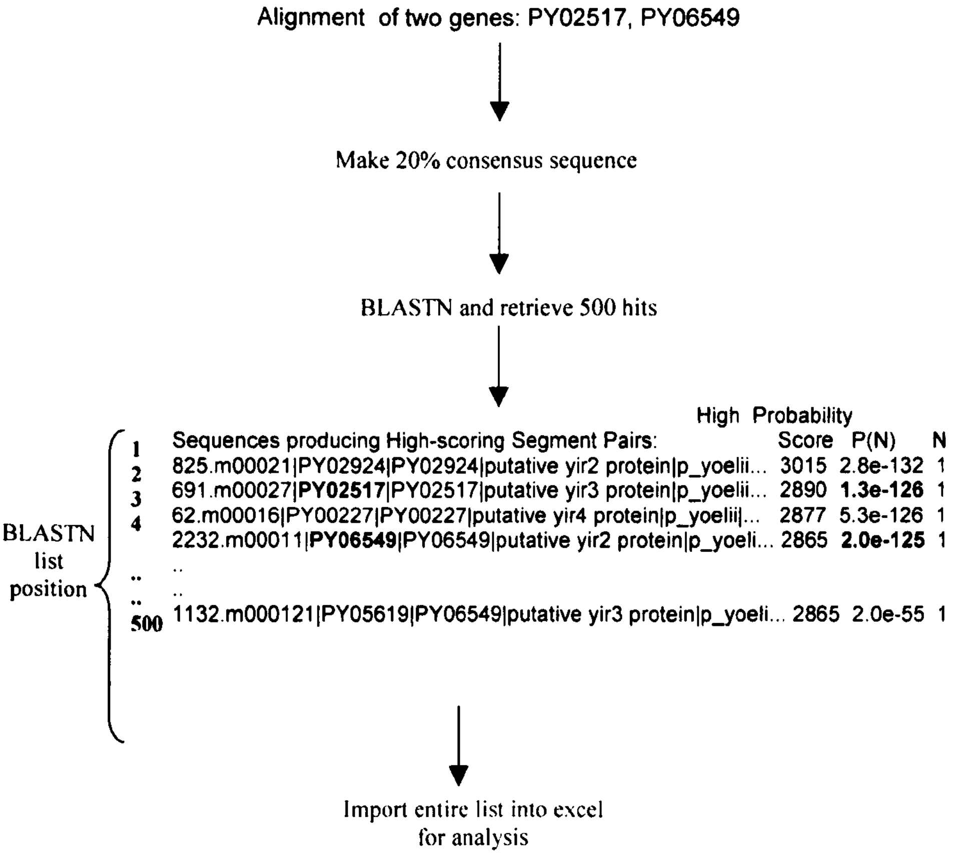

Percentage identical
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Alignment region

Figure 4.12

An example of how BLASTN was used for analysis

In this example, an alignment was made from two genes and a 20 % consensus sequence is generated from this alignment. The consensus sequence is then used to perform a BLASTN at www.tigr.org against the *Plasmodium yoelii* CDS database. The resulting BLASTN list in the example contains the genes in the alignment at position 2 and 4 respectively. The entire list is imported into excel where it is numbered 1 to 500. The positions of all self-hits are identified and are plotted in a graph with their position in the BLASTN list on the X-axis. The E value development for the entire 500 hits is plotted on a logarithmic graph alongside.



Alignment of two genes: PY02517, PY06549

Make 20% consensus sequence

BLASTN and retrieve 500 hits

```
                                                            High  Probability
1   Sequences producing High-scoring Segment Pairs:            Score  P(N)      N
2   825.m00021|PY02924|PY02924|putative yir2 protein|p_yoelii...  3015  2.8e-132  1
3   691.m00027|PY02517|PY02517|putative yir3 protein|p_yoelii...  2890  1.3e-126  1
4   62.m00016|PY00227|PY00227|putative yir4 protein|p_yoelii|...  2877  5.3e-126  1
    2232.m00011|PY06549|PY06549|putative yir2 protein|p_yoeli...  2865  2.0e-125  1
    ..
..  ..
..
500 1132.m000121|PY05619|PY06549|putative yir3 protein|p_yoeli... 2865  2.0e-55   1
```

Import entire list into excel
for analysis

An example (Fig. 4.12) of how a BLASTN analysis was used; the entire BLASTN list was retrieved and imported into Excel. Here the position of genes present in the alignment were identified and plotted in a graph, where the X-axis represent the position on the BLASTN list. Similarly, the E value development was used to create a log scale graph over the entire range of 500 hits, which was the maximum that could be retrieved at www.tigr.org. It was important to compare the position on the BLASTN list of self-hits from each of the supergroups and their E value development to genes that had not been phylogenetically grouped.

If any random alignment (i.e. not genes from any particular group/supergroup) of yir genes would pick up most of the genes in that alignment at the top of the BLASTN list with very low E values, it would reject the idea of superhomologues being identifiable through this method. Therefore, five sets of randomly chosen *yir* genes were aligned and consensus sequences generated from each of these alignments. There were 12, 14, 16, 60 and 50 genes in each of the five alignments. The latter alignment containing 50 *yir* genes, called "balanced random", was generated so it consisted of 10 genes from each of the five supergroups. The "balanced random" was made to check, if a different type of retrieval was observed when an equal number of genes from each of the supergroups were used as opposed to random picking, which is biased to lead to picking a large proportion of genes from the large SG1. From BLASTN with these consensus sequences (Fig. 4.13 a), it can be seen that the self-hits are scattered somewhat out over all the 500 hits on the BLASTN list. The e values increases rapidly within the approximately first 100 hits (Fig. 4.13 b) and then assumes a more steady development. Within these first 100 hits, the five test consensus sequences retrieved an average of 20% of the repertoire they were generated from. This has to be compared to the average 12% chance (100 out of 838)

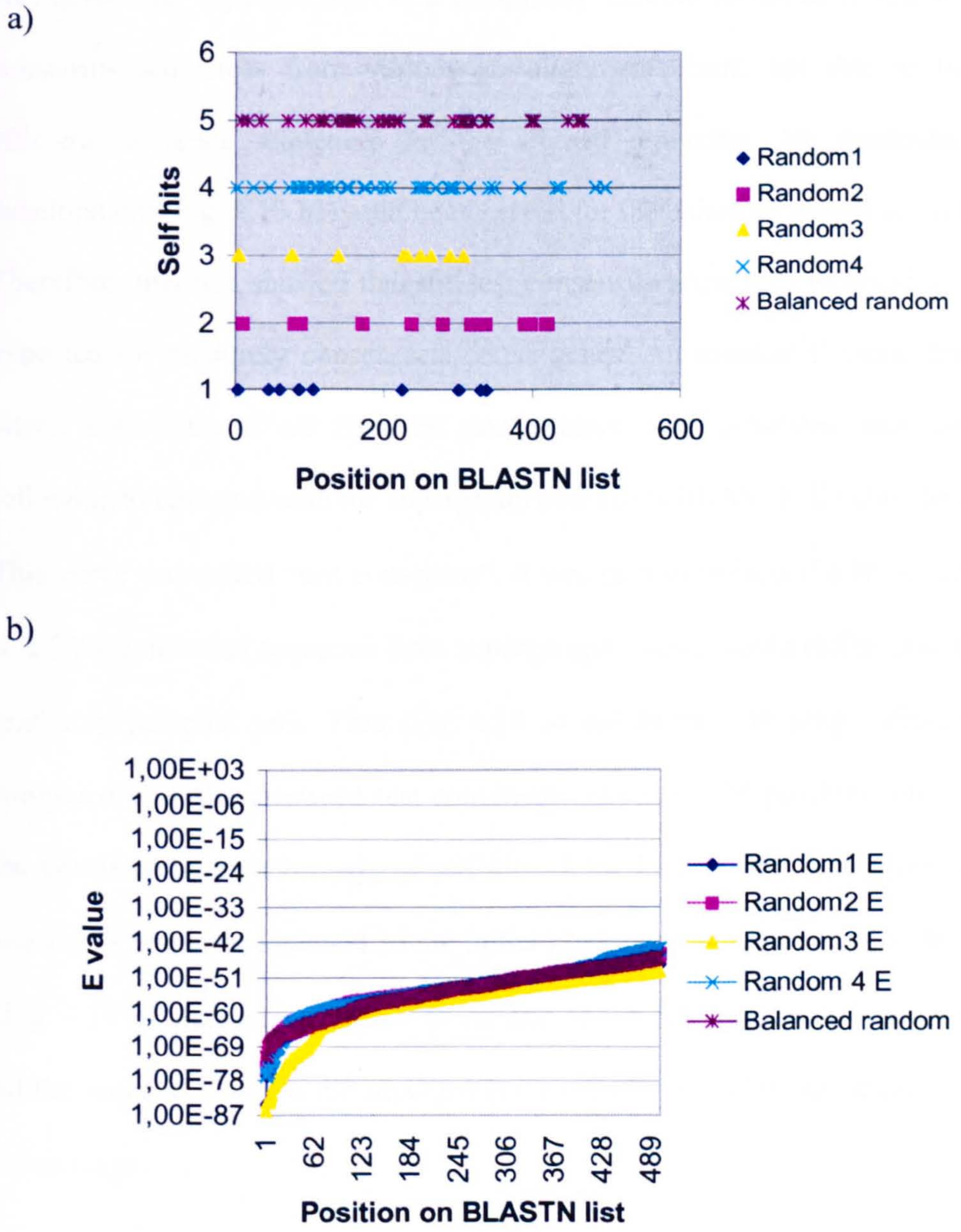
Figure 4.13

BLASTN with test consensus sequences

BLASTN analysis of 20% consensus sequences generated from five alignments containing the following numbers of *yir* genes: 12 (Random 1), 14 (Random 2), 16 (Random 3), 60 (Random 4), 50 (Balanced random). In each case, the BLASTN list was retrieved and used to investigate e value developments and self-hits.

a) Distribution of self-hits. On the X-axis is shown the position on the generated BLASTN list, while the Y-axis shows the self-hits occurring within each list for each of the five consensus sequences (1 to 5). Within the first 100 positions on the BLASTN list, the following percentages of self-hits had occurred for each of the five consensus sequences: Random 1: 33%, Random 2: 21%, Random 3: 25%, Random 4: 13%, Balanced random: 10%. This gives an average of 20% self-hits between the first 100 low scoring BLASTN list. This scattering of self-hits meant that the consensus sequences did not retrieve a high proportion of the aligned genes they were generated from.

b) The e-value developments for BLASTN with each of the five 20% consensus sequences. On the X-axis is shown the position on the BLASTN list, and on the logarithmic Y-axis is shown the corresponding E value for each hit. All the curves has a similar discontinuity with the approximately first 100 positions (on the X-axis) containing rapidly increasing E values, followed by a more steady E value development hereafter.


a)
6
5
4
3
2
1
Self hits
0
200
400
600
Position on BLASTN list
◆ Random1
■ Random2
▲ Random3
× Random4
✳ Balanced random
b)
1,00E+03
1,00E-06
1,00E-15
1,00E-24
1,00E-33
1,00E-42
1,00E-51
1,00E-60
1,00E-69
1,00E-78
1,00E-87
E value
1
62
123
184
245
306
367
428
489
Position on BLASTN list
Random1 E
Random2 E
Random3 E
Random 4 E
Balanced random

for any *yir* gene being randomly present within the first 100 hits. Although this is a bit higher than expected than in a completely random retrieval, it still showed that consensus sequences from various *yir* alignments were not able to function as efficient reference sequences for the aligned repertoire. No particular E value development (Fig. 4.13 b) could be observed for the balanced picked set in the graph. Therefore, this test showed that the test consensus sequences behaved as would be expected for randomly chosen sets of *yir* genes. An average E value development curve, consisting of all five test consensuses, was generated and used in the following to compare with the supergroup consensus BLASTN E value development. This curve was called "test consensus". It was then examined if a BLASTN analysis of a 20% consensus sequence from supergroup1 (SG1) would differ from that of the randomly sampled *yirs*. This (Fig 4.14 a) exhibited a striking difference, when compared with the averaged test consensus. The first 366 positions on the BLAST list consisted overwhelmingly of self-hits. Some hits could not be grouped, as their sequences were not included in the initial phylogenetic analysis. Also, the E values (Fig. 4.14 a) increased suddenly at the end of the self-hits and this could mean that all the sequences within the supergroup were within a highly statistical significant E value range.

What is clear from this is that SG1 can be defined by a range of E values, which are markedly lower than the test consensus. The same analysis was performed for the remainder of the supergroup consensus sequences SG2 to SG5 (Fig. 4.14 b to e). As can be seen in all cases, the supergroup in question came first on the BLASTN list. For SG2 (Fig. 4.14 b) and SG3 (Fig. 4.14 c), the E values were higher after the intersection point with the test consensus, whereas for SG4 (Fig. 4.14 d), the curves

Figure 4.14

BLASTN with 20% consensus sequences from the five supergroups

From alignments each of the five supergroups et al., 20% consensus sequences were generated and BLASTN analysis was performed individually. In the following figures, the X-axis shows the position on the BLASTN list; the left Y-axis shows hits within each of the five supergroups (1 to 5, marked as 🟥 in the graph). The right Y-axis shows the E values along the BLASTN list, and here the used supergroup consensus sequence for each particular supergroup is indicated as■ and the averaged test consensus sequences as🟥 .

a) BLASTN with a 20% SG1 consensus sequence. On the figure, a long continuous stretch of self-hits can be seen (1 on left Y-axis) until around position 331 on the x-axis. The SG1 e values intersect the test consensus e values shortly hereafter, at position 366.

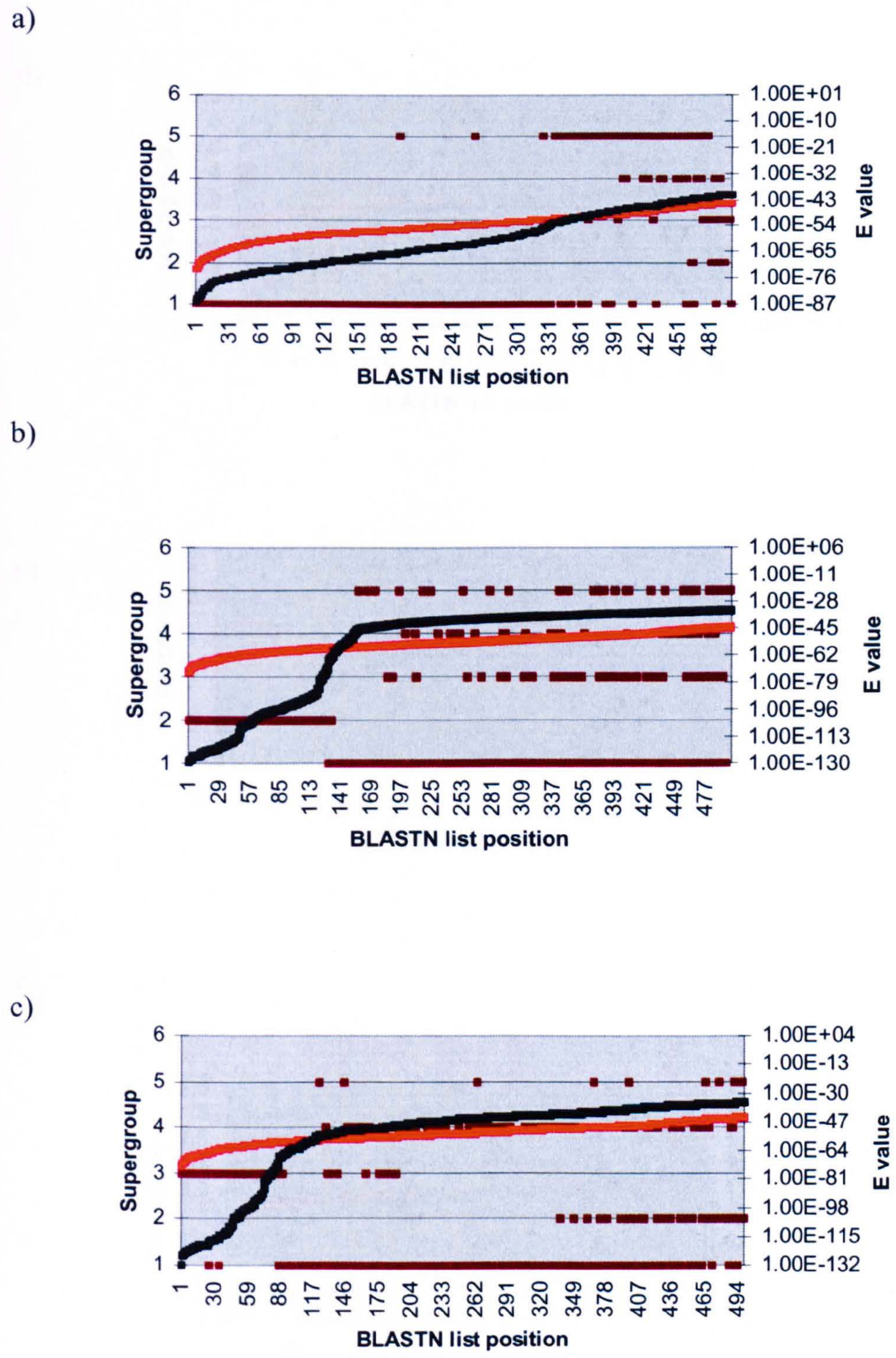
b) BLASTN with a 20% SG2 consensus sequence. The SG2 e values intersect the test consensus e values at position 140.

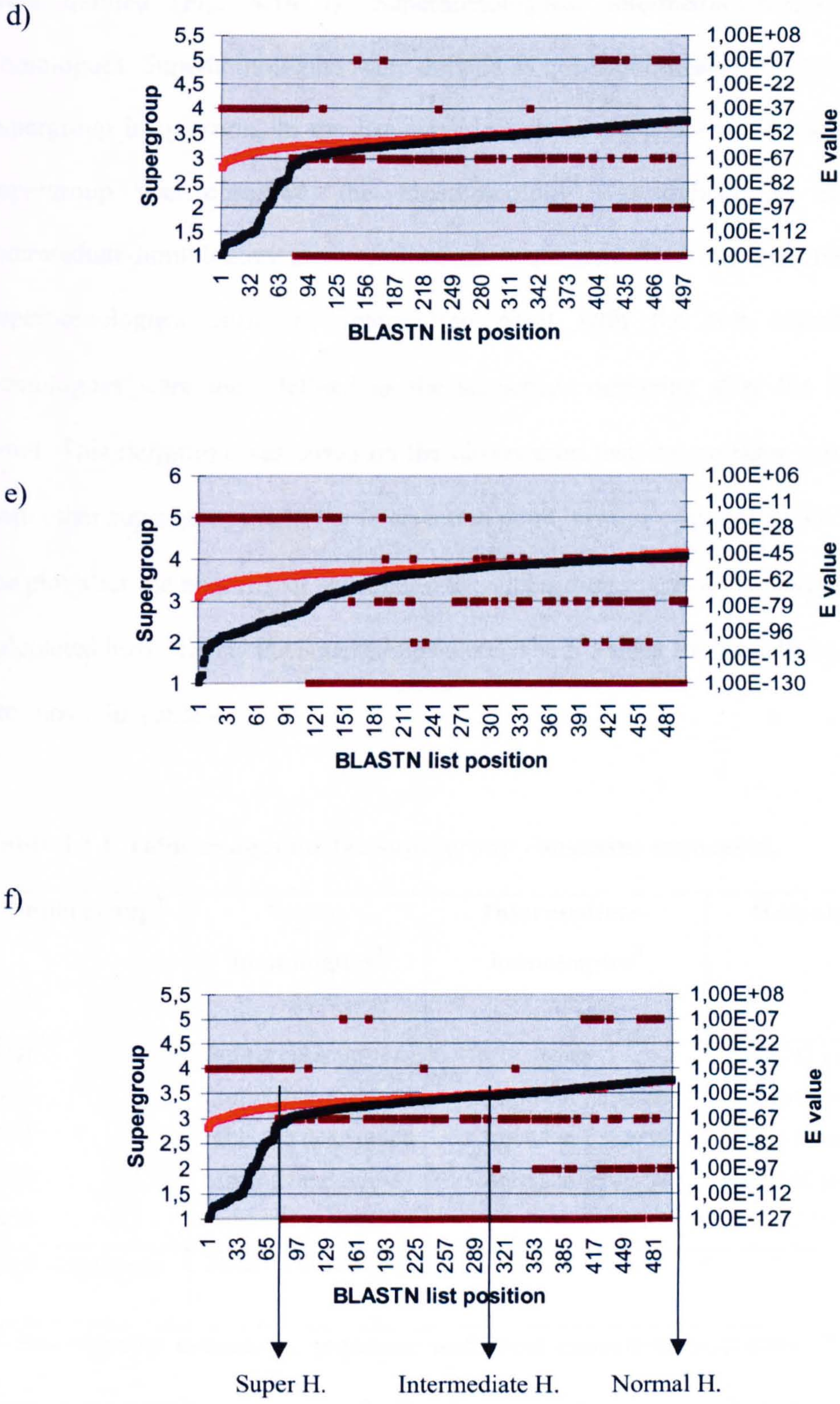
c) BLASTN with a 20% SG3 consensus sequence. The SG3 e values intersect the test consensus e values at position 120.

d) BLASTN with a 20% SG4 consensus sequence. The SG4 e values intersect the test consensus e values at position 396.

e) BLASTN with a 20% SG5 consensus sequence. The SG5 e values do not intersect with the test consensus e value development.

f) Graphical representation, exemplified by the SG4 BLASTN, of how superhomologues (Super H.), Intermediate homologues (Intermediate H.) and homologues (Normal H.) were defined.


a)
Supergroup
6
5
4
3
2
1
1
31
61
91
121
151
181
211
241
271
301
331
361
391
421
451
481
BLASTN list position
1.00E+01
1.00E-10
1.00E-21
1.00E-32
1.00E-43
1.00E-54
1.00E-65
1.00E-76
1.00E-87
E value
b)
Supergroup
6
5
4
3
2
1
1
29
57
85
113
141
169
197
225
253
281
309
337
365
393
421
449
477
BLASTN list position
1.00E+06
1.00E-11
1.00E-28
1.00E-45
1.00E-62
1.00E-79
1.00E-96
1.00E-113
1.00E-130
E value
c)
Supergroup
6
5
4
3
2
1
1
30
59
88
117
146
175
204
233
262
291
320
349
378
407
436
465
494
BLASTN list position
1.00E+04
1.00E-13
1.00E-30
1.00E-47
1.00E-64
1.00E-81
1.00E-98
1.00E-115
1.00E-132
E value


d)
Supergroup
5,5
5
4,5
4
3,5
3
2,5
2
1,5
1
E value
1,00E+08
1,00E-07
1,00E-22
1,00E-37
1,00E-52
1,00E-67
1,00E-82
1,00E-97
1,00E-112
1,00E-127
1 32 63 94 125 156 187 218 249 280 311 342 373 404 435 466 497
BLASTN list position
e)
Supergroup
6
5
4
3
2
1
E value
1,00E+06
1,00E-11
1,00E-28
1,00E-45
1,00E-62
1,00E-79
1,00E-96
1,00E-113
1,00E-130
1 31 61 91 121 151 181 211 241 271 301 331 361 391 421 451 481
BLASTN list position
f)
Supergroup
5,5
5
4,5
4
3,5
3
2,5
2
1,5
1
E value
1,00E+08
1,00E-07
1,00E-22
1,00E-37
1,00E-52
1,00E-67
1,00E-82
1,00E-97
1,00E-112
1,00E-127
1 33 65 97 129 161 193 225 257 289 321 353 385 417 449 481
BLASTN list position
Super H.
Intermediate H.
Normal H.

intersected at a much later point. For SG5 (Fig. 4.14 e), the two curves did not intersect at all. To analyse this in a comprehensible way, three levels of homology were defined (Fig. 4.14 f): Superhomologues, Intermediate-homologues and homologues. Superhomologues were defined as genes continuously belonging to the supergroup in question. In the list, when a gap of ten genes not belonging to the supergroup was observed, the superhomology was defined as ended. The intermediate-homologues were defined as sequences from the end point of the superhomologues until the intersection point with the test consensus. The homologues were then defined as the sequences occurring after the intersection point. This definition was based on the observation that the overlaps that did occur with other supergroups until the intersection point, always occurred in the last end of the plot after the majority of self-hits to the supergroup in question. Thus, it could be calculated how "clean" the supergroups were. The E values for these three categories are shown in Table 4.1.

Table 4.1 E value ranges for the supergroup consensus sequences.

Supergroup[1]	**Super-homologues**[2]	**Intermediate-homologues**[3]	**Homologues**[4]
SG1	6.80E-87 to 4.60E-51	None	4.60E-51 to 9.50E-42
SG2	6.30E-130 to 5.00E-64	8.20E-64 to 7.70E-59	1.20E-57 to 1.40E-34
SG3	2.80E-132 to 6.50E-68	8.70E-68 to 1.50E-58	3.70E-58 to 2.10E-36
SG4	1.30E-127 to 1.50E-67	5.10E-67 to 3.20E-50	4.10E-50 to 7.40E-45
SG5	6.30E-130 to 1.40E-76	1.60E-76 to 1.30E-47	None
Test consensus	2.81E-73 to 2.54E-45		

1: Supergroup consensus sequence and Test consensus sequence. 2: E value range for super-homologoues. 3: E value range for Intermediate-homologoues, 4: E value range for Homologoues. The entire E value range for the Test consensus sequence is also indicated.

Within the superhomologues range, the following percentages of the total repertoire of that SG were retrieved (%): SG1 96, SG2: 98, SG3: 89, SG4: 95 and SG5: 90. Together, this clearly showed that the consensus sequences from each of the supergroups did function as highly specific reference sequences, which were able to retrieve a significant proportion (89 to 96%) of the repertoire they were generated from.

4.2.14 Analysis of supergroups in *P. berghei* and *P. chabaudi*

Since *P.yoelii yir* genes have homologoues in the two other laboratory adapted rodent malaria species *P.berghei* and *P.chabaudi*, a comparative analysis was performed with the consensus sequences to see which of the *yir* supergroups had high scoring superhomologues in these two species. The strategy behind this analysis was a simultaneous BLASTN against genomic sequences at www.plasmodb.org for all three species. Using an approach similar to that used for the BLASTN analysis previously, the point where the E value curve changed trend dramatically was used to define where the super-homology ended. The assumption was that any sequence from the other two species that were able to replace any *yir* genes within that list would qualify as a super-homologue in those two species as well, and would indicate that this supergroup was also present in their genome.

From the results of this analysis (Fig. 4.15 a), it can be seen that intersections between the SG (1 to 5) consensuses and the test consensus occurred for each of the samples. When the number of hits before the intersection points (Fig. 4.15 b) were calculated, it is clear that only SG1 had hits in both *P.berghei* and *P.chabaudi*, while all SGs had hits to *P.berghei*. However, SG2 only had 5 hits to *P.berghei* compared to 18 for SG1, 18 for SG3, 14 for SG4 and 19 for SG5.

Figure 4.15

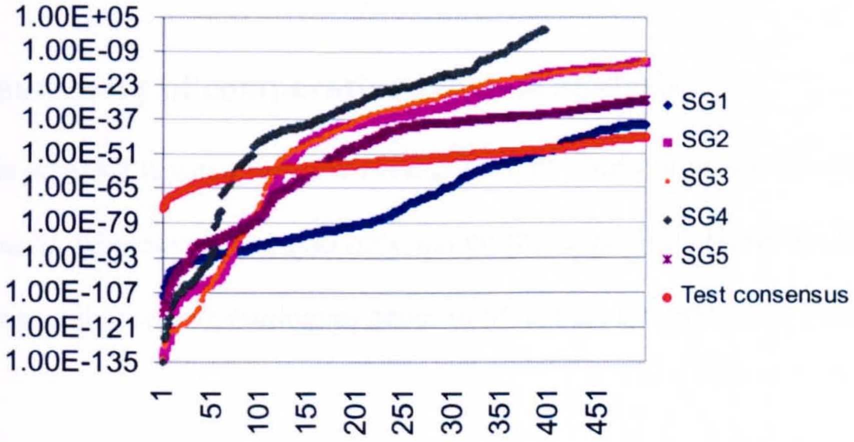
Comparative BLASTN

Retrieved hits from a BLASTN with the five supergroup consensus sequences against the mixed rodent genomic database at www.plasmodb.org were used to assess the presence of similar high-scoring superhomologues in *P. berghei* and *P. chabaudi*.

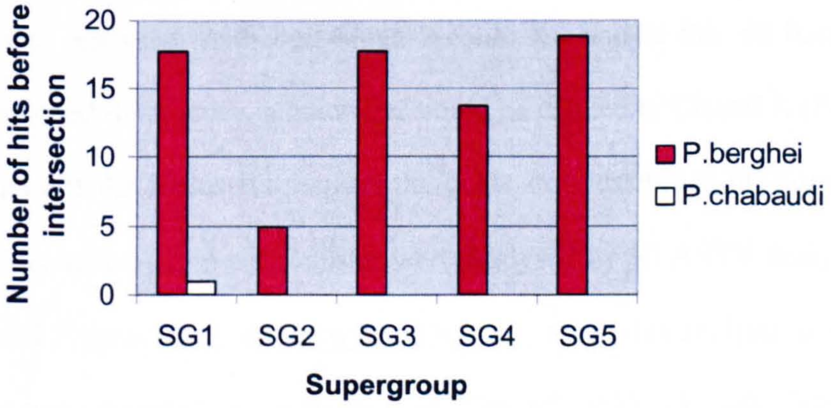
a) E value developments (left Y-axis) obtained from five individual BLASTN analyses with the five consensus sequences (SG1 to SG5, colour coded in key). The test consensus sequence was also included (see key), and in each case, the intersection point between this curve and the SG1 to SG5 consensus curves was used to define super-homologues.

b) Number of hits (Y-axis) to *P.berghei* and *P.chabaudi* genomic sequences with each of the supergroup's consensus sequences (SG1 to SG5, X-axis).

a)


1.00E+05
1.00E-09
1.00E-23
1.00E-37
1.00E-51
1.00E-65
1.00E-79
1.00E-93
1.00E-107
1.00E-121
1.00E-135
1
51
101
151
201
251
301
351
401
451
SG1
SG2
SG3
SG4
SG5
Test consensus

b)


Number of hits before intersection
20
15
10
5
0
SG1
SG2
SG3
SG4
SG5
Supergroup
P.berghei
P.chabaudi

In addition, the hits to *P.berghei* for SG2 occurred in the last parts of the curve (Fig. 4.15 a, at position 110, intersection was at position 123) were very close to the background, and therefore it could indicate that SG2 is specific for *P.yoelii* with no true super-homologues in the two other rodent malaria species. The first *P.chabaudi* SG1 hit was significant and occurred at position 203 at an E value of 5.61^{-56}.

4.2.15 Summary of comparative genomic analysis

From this analysis it was clear that *P.berghei* share four out of the five supergroups with *P.yoelii*, whereas *P.chabaudi* only shared the large SG1. However, the analysis also indicated that superhomologues genes to SG2 were not present in *P.berghei.*

4.2.16 Characterisation of the remaining *yir* repertoire

Some 120 annotated *yir* genes had been removed from the phylogenetic analysis either because they could not be translated or because they were too divergent for phylogenetic analysis. Towards the end of this project, it was discovered that among these 120 genes there were a set of very large annotated *yir* genes (identified by PhD student Sandra Koernig). To investigate this further, the remainder of the *yir* repertoire was analysed. Although Mega 2 could not handle this set because of its before-mentioned divergence, a guide tree could be created in Clustal X (Fig. 4.16 a). To investigate if the 8 clusters seen on the guide tree resembled the supergroups, a number of sequences from each cluster were analysed by BLASTN analysis against the annotated *P.yoelii* genes at: www.http://tigr.org. From this analysis, the following similarities were established: cluster1: SG3, cluster2: SG5, cluster3: SG5, cluster4: no high scoring matches, cluster5: SG2, cluster6: SG1, cluster7: SG1, cluster8: SG1. Thus, seven out of the eight clusters bore resemblances to the already established five supergroups.

Figure 4.16

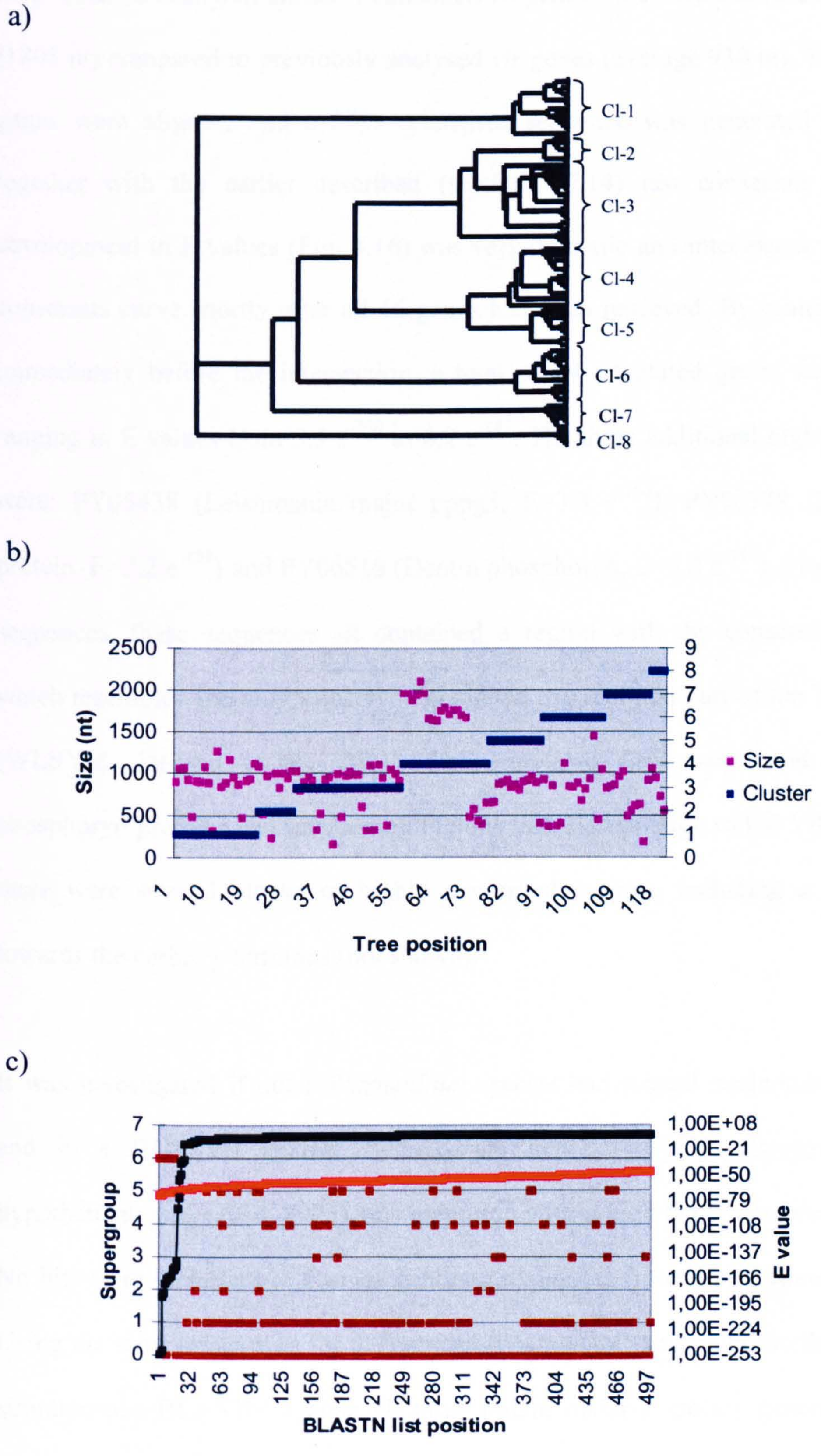
Analysis of the remaining *yir* repertoire

The remaining 120 *yir* genes, which could not be analysed by previously used phylogenetic methods, were analysed. Initially, a multiple alignment of these 120 full-length sequences were made in ClustalX. From this, a guide-tree was constructed and analysed further.

a) Guide tree of the 120 *yir* genes. On this tree, eight clusters were observed (Cl.- 1 to Cl. -8).

b) While keeping the eight clusters (from a) on the X-axis, the size distribution of the genes in each cluster is plotted on the Y-axis. This revealed that cluster 4 contained the very large *yir* genes with an average length of 1805 nt.

c) BLASTN with the cluster 4 *yir* genes. On the X-axis, the position on the BLASTN list, on the left Y-axis: supergroups 1 to 5 and 6 is the large cluster 4 *yir* genes, right Y-axis: E values. The test consensus sequence E values were: 🟧 , the cluster 4 E values were: ■ , while the supergroups (1-5) and large *yir* self-hits were: 🟥 . It can be seen that self-hits occur at extremely low e values before hits to any of the supergroups occur.


a)
Cl-1
Cl-2
Cl-3
Cl-4
Cl-5
Cl-6
Cl-7
Cl-8
b)
Size (nt)
2500
2000
1500
1000
500
0
9
8
7
6
5
4
3
2
1
0
Size
Cluster
1
10
19
28
37
46
55
64
73
82
91
100
109
118
Tree position
c)
Supergroup
7
6
5
4
3
2
1
0
1,00E+08
1,00E-21
1,00E-50
1,00E-79
1,00E-108
1,00E-137
1,00E-166
1,00E-195
1,00E-224
1,00E-253
E value
1
32
63
94
125
156
187
218
249
280
311
342
373
404
435
466
497
BLASTN list position

The sizes of the genes in these eight clusters were also investigated (Fig. 4.16 b), and here it can be seen that cluster 4 contained 16 genes which were considerably longer (1805 nt) compared to previously analysed *yir* genes (average 930 nt). These sixteen genes were aligned, and a 20% consensus sequence was generated and plotted together with the earlier described (section 4.2.14) test consensus curve. The development in E values (Fig. 4.16) was very dramatic and intersected with the test consensus curve shortly after all 16 genes had been retrieved. By omitting two hits immediately before the intersection, a total of 19 annotated genes were retrieved ranging in E values from 5.5 e^{-250} to 6.2 e^{-147}. The three additional high-scoring hits were: PY05438 (Leishmania major pppg3, E=3.3 e^{-166}), PY00748 (Hypothetical protein, E=2.2 e^{-151}) and PY06516 (Dentin phosphoryn, E=6.2 e^{-147}). From translated sequences, these sequences all contained a region with the consensus WLSNK, which resembled the most conserved predicted extracellular part of the YIR proteins (WLSYM, Janssen et al., 2004). The annotated Leishmania and the Dentin phosphoryn proteins had intervening regions that did not align to the YIRs; however there were several blocks of highly conserved regions, including a large block towards the carboxy terminus (not shown).

It was investigated if other *Plasmodium* species had related nucleotide sequences, and in a BLASTN against *P.falciparum* genes (www.http://plasmodb.org), a hypothetical gene (*Pf*14_0023) was identified with a high E value score (E=1.0 e^{-6}). No hits were identified in *P.vivax* genomic sequences (incomplete coverage so far). Using the same strategy as for the comparative genomic analysis described above by simultaneous BLASTN against all three rodent malaria species genome, two hits (Contigs: Pb5456 and Pb5483, E= 7.9e^{-160} and 5.4e^{-218}) were obtained in *P.berghei.*

These were within the E value range (from 5.5 e^{-250} to 6.2 e^{-147}) for the 19 genes described above.

4.3 Summary and discussion

YIR phylogeny

In this chapter the organization of the *yir* genes was investigated. From an NJ tree showing both distance and topology, ten individual groups were identified. These ten groups were also consistently identified on trees generated with the ME and MP methods, and the differences between the trees, all occurred within the five major branch points. These branch points were therefore classified as five supergroups (SG1 to SG5), and the bootstrap values supporting this ranged from 51% for SG1, 86% for SG5 and 95% for SG2, SG3 and SG4.

Reliability of *yir* supergroups.

The reliability of a tree based on its bootstrap value can be defined as "the probability that a member of a clade is always a member of that clade", and in very general terms, bootstrap values below 25% are not reliable, whereas values above 90% can be considered reliable (Hall et al., 2004, 2nd edition). For example, in an analysis of the *expansin* multigene family in different plant species and using a NJ tree building method, bootstrap values above 60% were considered, and values above 80% were considered highly significant (Li et al., 2002). In another study, analysing the *hedgehog* and *hox* gene phylogenies in different species of fish with MP and NJ methods, branches with less that 50% bootstrap values were collapsed into polytomies (Zardoya et al., 1996). Similarly, in an analysis of the olfactory receptor proteins in various taxonomic groups using the NJ method, bootstrap values above 80% were considered to support a branch (Niimura et al., 2005). However, it should be noted that the above mentioned studies all analysed sequences across taxa, and this could mean that more diversity is present in these datasets than in the present. Nevertheless, these earlier used bootstrap values were considered as a guideline in this analysis, and by these criteria, SG2 to SG5 (bootstrap values from 86 to 95%) were considered reliable supergroups. The weak phylogeny for SG1 (bootstrap value of 51%) suggested that this supergroup consisted of a more loose set of related genes. The BLASTN showed that, in accordance with the phylogeny, the superhomologues within SG1 were supported by higher E values (i.e. less significant) than SG2 to SG5. Despite this, the BLASTN also showed that the SG1 genes were more related to each other than to any random set of *yir* genes, or to genes from SG2 to SG4. Therefore, SG1 was considered a weak supergroup.

Comparison of *yir* sequences to TIGR annotation

When the supergrouping was compared to the TIGR annotation groups (Carlton et al., 2002) it was found that one of the supergroups (SG2) especially contained a mixture of the groups proposed by TIGR. This strongly suggested that this supergroup was not identified in the initial analysis, because of the absence of a suitable reference sequence, and thus this study, performed in this thesis, is an improvement of the previous *yir* classification.

Size and GC content of genes in supergroups

The *yir* genes in the different groups could be divided into two distinct size categories; one consisting of genes between 773 and 825 bp (and consisting of groups 1, 2 , 4, 5 and 6) and one consisting of genes between 884 and 924 bp (consisting of groups 3, 7, 8, 9 and 10). The standard size deviations within the groups were between 36 and 58 bp, and this meant that there were between 1 and 4 standard deviations between the two sets. Group 7-8 comprised SG4 and Group 9-10 SG5, and these genes were very similar in sizes. This was not the case for the groups in SG1 (773-825 bp), since Group 3 consisted of much larger genes (average 908 bp). This indicated that size alone had not determined the phylogeny, since Group 3 was not located together with the similar sized SG4 or SG5 genes. Considering that SG1 was supported by the lowest bootstrap values, this is not surprising, and indicates that SG1 is more heterogeneous than any of the other supergroups. On the other hand, SG4 and SG5 were supported by high bootstrap values, and also consisted of similar sized genes. This finding therefore supports the classification of *yir* genes into five different supergroups.

A highly heterogeneous GC distribution was found for all groups, and the maximum difference in GC content was 1.98% (Group 4: 25.01% and Group 8: 26.99%). The standard percentage variation within each group was between 0.7 and 1.27% (Group 7 and 5 respectively), and this meant that there were between 1.6 and 2.8 standard deviations between the two most different groups. However, despite some distinct differences in GC content between the groups, it was not possible to assign these to any pattern reflecting the supergroups or bootstrap values from the phylogenetic trees. Another NJ tree (not shown) was constructed by using the Tamura-3 parameter, which takes GC bias into account. However, the groups clustered similarly on this tree, and their order of appearance on the tree did not reflect their average GC contents. Therefore, the GC content had not been a defining factor for the groups/supergroups. It is possible that the GC content signifies predominant location of certain groups on distinct chromosomes, but this is an open question.

Localization of *yir* genes

No *yir* genes co-localized with annotated housekeeping genes, known from synteny analysis to be located in chromosomal internal regions (Carlton et al., 2002 and 2005). Telomere-associated repetitive elements (TARE 1-6) exist in *P.falciparum* but are not similar to those in *P.yoelii* (Figueiredo et al., 2000 and Carlton et al., 2002). Nevertheless, several repeat elements were found on several *P.yoelii* contigs and these were classified as subtelomeric contigs (Carlton et al., 2002). Location of the supergroups on these subtelomeric contigs (along with the 28 annotated telomeric contigs), revealed a striking distribution. Supergroups SG1 to SG5 all had a distinct proportion of their genes located on these contigs.

Unfortunately, the lack of contiguous sequences from the chromosome ends, makes this analysis a snapshot of how the supergroups are distributed at a specified location. It was interesting to note that G3, a component group of SG1, had a relatively higher proportion of genes located on the subtelomeric contigs, and this also correlated with genes in this group being more comparable in sizes to the highly subtelomerically located SG4 and SG5 genes. This suggests that the reason for the low bootstrap value for SG1 could be that it is a mixed supergroup, which co-localizes with several other supergroups.

It could be that SG1 genes are located on several chromosomes and in several chromosomal regions, whereas the other supergroups could have a more limited distribution. This finding also supports the phylogenetic classification of *yir* genes.

In this study, there were three lines of evidence supporting the proposed *yir* grouping:

1. Identical supergrouping in three tree building models, and high bootstrap support for four of the five supergroups
2. Supergroups with a homogenous size distribution of their constituent groups (e.g. SG4 and SG5) was supported by higher bootstrap values than SG1, which had a very heterogeneous size distribution.
3. There was a clear correlation between the five supergroups and their proportionate location on subtelomerically annotated contigs.

In *P.falciparum*, the *var* genes can be divided into five groups (upsA to upsC and upsB/A and ups B/C) based on their 5′ intergenic regions, and this correlates with their chromosomal localisation (Gardner et al., 2002 and Lavstsen et al., 2003). The *var* genes within each group encode distinct adhesive domains (Gardner et al., 2002 and Lavstsen et al., 2003). Although shuffling of domains is possible, certain domain combinations are not observed, probably because they are not biologically favourable, and they are therefore not selected (Gardner et al., 2002 and Kraemer et al., 2003). The selection of certain groups of genes encoding distinct domains could have emerged through localization of these groups in distinct chromosomal regions where mixing with genes encoding the unwanted domains is not favoured. Therefore, the grouping and localization of *var* genes appears to be very much linked to the maintenance of genes encoding proteins with distinct functions. This study found that the *yir* genes are organised into five supergroups, similar to *var* genes. One of these supergroups appeared to be more heterogeneous than the remaining four. The first two lines of evidence for the proposed *yir* grouping both rely on the actual sequences, the third does not. Secondarily, compartmentalisation of the *yir* repertoire would be a highly efficient way of maintaining discrete supergroups, which could perform different functions and also be regulated differentially. This is not unprecedented, as this was, as discussed above, also found to be the case for the *var* genes. Towards the end of this PhD thesis, a study of the *vir* genes of *P.vivax* was published (Fernandez et al., 2005), showing that the *vir* repertoire is organized in five highly diverse groups with different size distributions. It had been suggested earlier that VIR proteins encoded by a highly diverse set of genes could be involved in antigenic variation, whereas more conserved sets could be involved in adherence to barrier cells in the spleen (del Portillo et al., 2004). If this is the case for the YIR proteins, SG1 would be the prime candidate for antigenic variants, whereas SG2 to SG5 could be involved in specific functions relying on a higher degree of conservation. However, it must be said, that the differences between the YIR proteins from the different supergroups are not as large as the completely different architecture of the Pfemp encoding domains of the *var* groups (Gardner et al., 2002).

Comparative analysis of *yir/bir/cir* organisation

The comparative analysis of the *bir* and *cir* grouping in *P.berghei* and *P.chabaudi* respectively indicated that there were *bir* superhomologues to four out of the five supergroups, while the *P.chabaudi cir* genes only had superhomologues to SG1. In other phylogenetic studies, based on *cythochrome c* and *dhfr* genes (Perkins et al., 2002 and Rich et al., 1998), it was found that *P.berghei* resembles *P.yoelii* more than does *P.chabaudi*. With the current status of the *P.berghei* sequencing, 180-245 *bir* genes have been estimated (preliminary Sanger Center annotation data). This number might change somewhat once the sequencing coverage (currently around 3.5 X coverage) is increased; however it is very unlikely that there are as many *bir* as *yir* genes. The analysis described in this thesis shows that both *P.berghei* and *P.yoelii* have retained all supergroups, except SG2, from their last common ancestor. In *P.yoelii*, there are more genes within these shared supergroups (597) than the highest estimate of *bir* genes (245). The most likely explanation for this is a higher rate of gene duplications in *P.yoelii* than *P.berghei*. It cannot be excluded that *P.berghei* contains additional groups of genes that differ from *P.yoelii*, as no phylogeny has been made for the *bir* genes. Since SG2 was found to be specific for *P.yoelii*, these genes have either evolved independently in *P.yoelii* since speciation, or have been removed through selection from *P.berghei*. Therefore, this analysis shows that two different evolutionary mechanisms have formed the *bir* and *yir* genes: duplication (or reduction) of an existing repertoire in SG1 and SG3 to SG5, and possible *de novo* generation (or complete removal) of the SG2 repertoire. Although genetic drift could have removed a low copy number of SG2 superhomologues from modern *P.berghei*, there must have been a positive selection for SG2 genes in modern *P.yoelii*, as this is the second largest of the supergroups with 121 genes.

Large *yir* genes

The large *yir* genes had very little similarity to the remaining *yir* repertoire, and were for that reason excluded from the phylogenetic analysis. They comprised a distinct set of superhomologues, also including three genes annotated as hypothetical, Leishmania major pppg3 and Dentin phosphoryn. They were very distantly related to the remaining *yir* repertoire as determined by e value comparisons. They had probably been included in the *yir* repertoire because they contained the WLS amino acid motif, found to be highly conserved in PIR proteins (Janssen et al., 2004). In *P.chabaudi*, ten multigene families have been found in subtelomeric regions, and some of these resembled each other in gene structure (Fischer et al., 2003). It is plausible that several other multigene families are also present in *P.yoelii*, and some of these could resemble the *yir* genes. Whether this (weak) similarity also means shared function is an open but highly important question.

Chapter V

Transcription profile

5.1 Introduction

Studies of the *var* genes in *P.falciparum* have found that they are expressed in a sequential and mutually exclusive manner (Fernandez et al., 2002 and Scherf et al., 1998 and Chen et al., 1998). The switch rate for the *var* genes has been estimated to be 2.4% per generation in the absence of immune pressure (Roberts et al., 1992) and it has been proposed that *var* genes have distinct on and off rates (Horrocks et al., 2004). This would imply that there is a hierarchy of *var* genes likely to be transcribed and translated. Together these studies suggest that *var* genes are regulated in such a way that only a limited number of the Pfemp proteins are exposed to the host's immune system at the same time. This ensures, that during the course of the infection the parasite is able to express new variants of Pfemp that have not previously been seen by the immune system. This allows the parasite to avoid host immunity for a longer period, thus enhancing the likelihood of transmission.

An interesting question is whether a similar strategy would be employed when the total number of variant genes is significantly increased from 60 members to 838 as in the case of the *yir* family. This large repertoire of variant genes may allow the parasite to simultaneously transcribe and express a large number of variants thereby overwhelming the host immune system. An effective immune response to some variants would under these circumstances only affect a relatively small proportion of all parasites. Therefore, it was important to investigate how many genes are transcribed in the blood stages, as a high number of transcripts would seem to support the notion that *yir* genes utilize a different immunological strategy than the *var* genes. The five supergroups, described in the last chapter, may represent YIR that are expressed in different parasite life stages. The supergroups contain genes/proteins that are the most different within the *yir*/YIR repertoire. It would be reasonable to assume that if these differences are immunologically important, an antibody against a supergroup 1 YIR would be more likely to cross react with other proteins from the same supergroup 1 while being less efficient against YIR from another supergroup. Transcription and subsequent translation of *yir* from all the supergroups in the bloodstage, would therefore represent the most diverse repertoire of potential epitopes.

If indeed *yir* from all the phylogenetic groups were identified it would indicate that *P. yoelii* has the capacity to simultaneously express proteins representing the highest level of diversity within its repertoire. To ensure that in this study the effect of host immunity on the transcription of *yir* was minimized, all parasite samples were grown in Rag 2 -/- mice. These mice lack the ability to produce T and B cells through a deletion in the *recombinase II* gene, necessary for re-arrangement of T and B cell receptors (Shinkai et al., 1992).

5.1.1 Objectives

It was investigated through RT-PCR how many transcripts were present in a population of parasites in the absence of T and B cells. This was obtained through the optimisation and design of primers for RT-PCR, which would amplify cDNA specifically (5.2.1), as genomic contamination was a problem given the *yir* copy number. It was investigated if *yir* mRNA could be detected in all stages (5.2.2). Primers were designed which could potentially amplify transcripts from all supergroups (5.2.3). A large number of clones were analysed through sequencing (5.2.4) and the performances of the different primers were evaluated (5.2.5), followed by the identification of the extent of the transcribed *yir* repertoire (5.2.6).

5.2 Results

5.2.1 Avoiding genomic contamination

In order to avoid genomic contamination, a dual strategy was employed: 1.) to reduce the level of genomic DNA as much as possible during the RNA extraction steps. 2.) to use primers that would not amplify genomic DNA. During the initial RNA extraction step, using TRIzol, great care was taken to avoid the interphase layer, where the bulk of genomic DNA would be located. After this extraction, the RNA was digested with DNase I for 1 hour and re-extracted using phenol/chloroform (pH 4.2), again avoiding the interphase layer. Although this approach tended to reduce the level of detectable genomic contamination, it led to a significant loss of starting RNA material. Furthermore, analysis of the RNA by gel electrophoresis indicated partial RNA degradation. Therefore, this step was omitted and instead great care was taken to avoid the interphase layer during the initial TRIzol extraction step, to reduce contamination with genomic DNA as much as possible. For this reason another strategy to avoid amplification of genomic DNA was used. Using primer sets where one primer was spanning the highly conserved exon 2/exon 3 splice site boundary (Fig. 5.1 a, set F and G1-G6), which would be expected only to amplify correctly spliced cDNA.

Initial tests using both splice-site spanning as well as primers located within exon 3 (Fig. 5.1 b) showed that the splice site spanning primers were unable to amplify a specific product from gDNA (Fig 5.1 b, Set F on dilutions). In contrast, primers located solely within exon3 gave the expected PCR product when using gDNA (Fig 5.1 b, Set C on dilutions).

Figure 5.1

Test of primers and RT-PCR analysis

a) Location of the used primer pairs. For the different primer sets, the expected size ranges of products were: Set C: 750-850 bp (gDNA) and 660-760 bp (cDNA), Set F: 650-750 bp, Set G1-G5: 500-600 bp, Set G6: 650-750 bp.

b) PCR was performed with primer set C on diluted cells (d.1: 100 cells, d.2 10 cells, d.3 1 cell). The products from this reaction were approximately 780 bp, which was within the expected size range. PCR with primer set F was performed on the same dilution, and no bands were visible. However, when primer set F was used on two cDNA samples (cDNA1 and cDNA 2, +/- RT enzyme), bands of around 750 bp could be seen. A 100 bp marker (M1) was also loaded on the gel for size comparisons. Nothing was seen in the –RT or Control (C.) lanes.

c) PCR with primer set F on +/- RT samples constructed from parasites separated into Schizonts (Sz), Late Trophozoites (L-Tz), Middle Trophozoites (M-Tz), Early Trophozoites (E-Tz) and Rings (R) (see materials and methods). The products led to the identification of bands above 700 bp in size when compared with the 100 bp (M1) size marker.

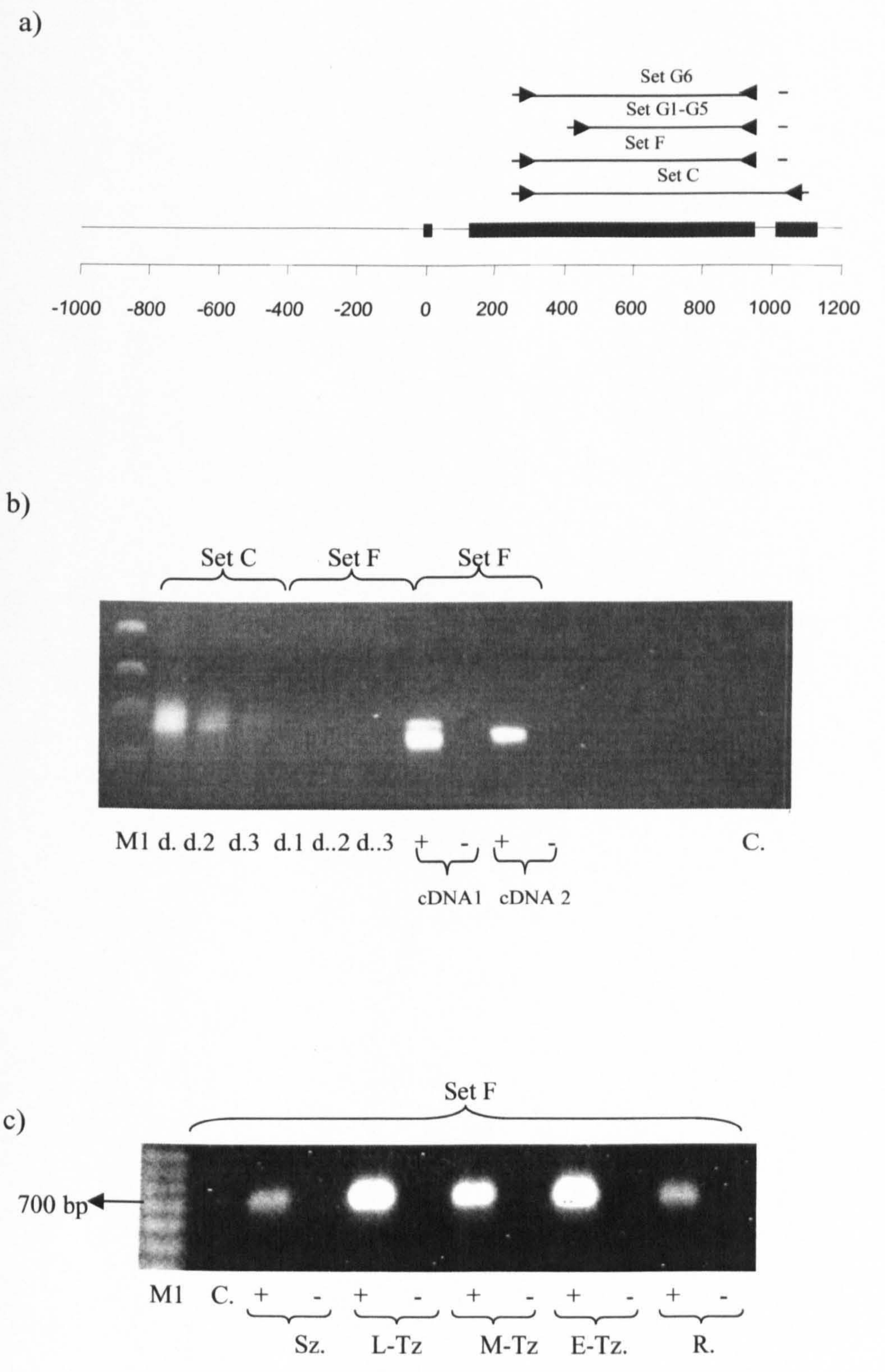
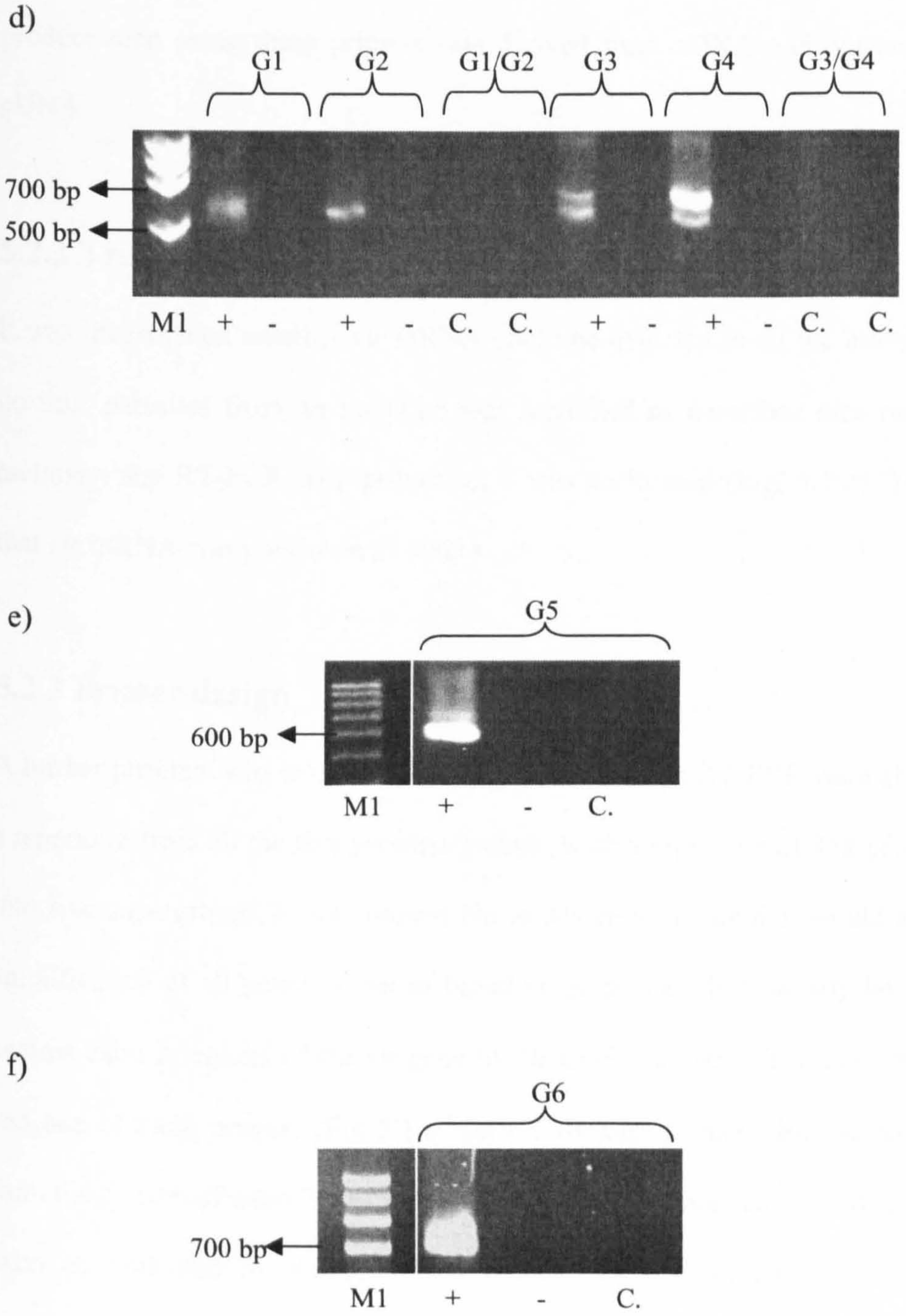

a)
Set G6
Set G1-G5
Set F
Set C
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
Set C
Set F
Set F
M1 d. d.2 d.3 d.1 d..2 d..3
+ - + -
cDNA1 cDNA 2
C.
c)
Set F
700 bp
M1 C. + - + - + - + - + -
Sz. L-Tz M-Tz E-Tz. R.

Figure 5.1..continued

d) PCR on cDNA with primer sets Gl to G4. Bands of around 600 bp can be found in all +RT lanes when comparing to the 100 bp size marker (M1) whereas nothing is seen in the –RT or Control (C.) lanes. For both primer sets G3 and G4, two differently sized transcript populations can be seen.

e) PCR on cDNA with primer set G5. A band of around 600 bp in size can be seen whereas nothing is seen in the –RT or Control (C.) lanes. A somewhat smaller intensity band can be seen above the major band.

f) PCR on cDNA with primer set G6. A band of around 700 bp in size can be seen whereas nothing is seen in the –RT or Control (C.) lanes.


d)
G1
G2
G1/G2
G3
G4
G3/G4
700 bp
500 bp
M1 + - + - C. C. + - + - C. C.
e)
G5
600 bp
M1 + - C.
f)
G6
700 bp
M1 + - C.

On the other hand, the splice site containing primers efficiently amplified up the correct sized (around 750 bp) product when using cDNA expected to contain the spliced product (Fig 5.1 b, Set F on cDNA 1 and cDNA 2). This confirmed that the splice-site spanning primer was able to selectively amplify cDNA, generated from spliced RNA. All reverse transcription experiments carried out in this chapter, included a –RT (no Reverse transcriptase added) control. This ensured that any product seen using these primers was derived from cDNA and not contaminating gDNA.

5.2.2 Transcription in stage separated parasites

It was investigated whether *yir* mRNA could be detected in all the blood stages. To do this, parasites from an infection was separated as described (see materials and methods) and RT-PCR with primer set F was performed (Fig. 5.1 c). This showed that *yir* mRNA was present in all stages.

5.2.3 Primer design

A further problem was to ensure that the primers used in RT-PCR were able to detect a repertoire from all the five *yir* supergroups. With a repertoire of 838 genes, divided into five supergroups, it was impossible to design a primer that would allow equal amplification of all genes. A set of broad range primers had already been designed against exon 2 regions of the *yir* gene by Dr Chris Janssen, University of Glasgow, and one of these primers (Fig 5.1 a, Set G1 forward primer), located some 300 bp from the 5′ start of exon 2 was found to be against the most conserved region in this exon as evaluated by the consensus function of a complete *yir* alignment. By manually examining this region in individual alignments of the supergroups, four additional primers (Fig 5.1 a, G2 to G5 forward primers) were designed from this region. One additional primer located in a conserved region found in the SG4 alignment, some 150 bp from the 5 ′ start of exon 2 was also designed (Fig 5.1 a, G6). A BLASTN was performed with each of these primers to see how many hits in total and full/nearly full matches they would produce (Table 5.1).

Table 5.1. Forward *yir* primer sequences.

Primer[1]	Sequence[2]	BLASTN hits[3]
G1	ATATGGTTAAGTTATATGTTAAACC	373
G2	ATGGTTAAGTTACAAATTAAACCAAAA	174
G3	ATTATGATATGGTTAAGTTATAAATTAA	436
G4	ACATTATCATATGGTTAAGTTATAAACTA	450
G5	GCTATTTTATGGTTAAGTTATAAAC	107
G6	GATAAGATTAATGCTGGATGTTTATG	151

All primers are forward primers used in the described primer sets (Gl to G6). 1: Primer set. 2: Forward primer sequence. 3: BLASTN hits obtained from www.tigr.org.

The BLASTN repertoire contained progressively poorer matches between the primer and sequences in the database. To establish which matches could be considered to have significant specificity to the primer, the BLASTN outputs were evaluated both in Bit Score and E values. The E values were found to develop very little on the BLASTN list, as the input sequences were very small, whereas the Bit Score development changed more systematically. Only the two top Bit Score values were used to estimate the specificity of the primers. When looking at the actual degree of matching within these two top Bit Scores it was found that on average they corresponded to one nucleotide mismatch between the primer and sequences in the database. Within this group a maximum of one mismatch between primer and sequences in the database was observed. The top hits for each primer were then classified based on the supergroup they hit (Table 5.2).

Table 5.2. *Yir* primer specificity.

Primer[1]	Top hits[2]	SG1[3]	SG2[3]	SG3[3]	SG4[3]	SG5[3]
G1	129	92	0	0	1	7
G2	27	0	0	100	0	0
G3	28	3	0	0	79	18
G4	26	0	0	0	0	100
G5	84	0	98	0	0	2
G6	25	4	0	0	84	12
Splice-site	414	51	74	37	41	49

1: Primer name (G1 to G6 forward primers and the reverse splice site primer). 2: Number of top hits on the BLASTN lists. 3: Percentages of top hits belonging to the different supergroups (SG1 to SG5). Note that for the splice site primer, numbers are percentages of each supergroup.

It is clear from Table 5.2 that each of the six primers had their best matches with genes from the supergroups they were designed against. However, for all primers there were several other, lower scoring hits from the BLASTN list. As Table 5.2 only shows the top hits (i.e. a maximum of 1 nt mismatch over the entire aligned range), the primers could potentially amplify transcripts from outside the supergroup they were designed for. The repertoire of the exon 2/exon 3 splice site primer was also investigated through BLASTN, and it was found that the hits were distributed as follows: 117 perfect matches, 192 with one mismatch and 105 with more than one mismatch, and for the worst of those 4 nucleotides did not match to the sequence. These numbers most likely represent a huge underestimation due to the misannotations of a large number of genes, where the second and third exons were not joined (see chapter III). In addition, the splice site primer was not biased towards any of the supergroups. RT-PCR with these primers was performed on reverse transcribed RNA from a blood stage sample grown in immunodeficient Rag 2 -/- mice (see materials and methods). Initially, the PCR conditions were determined empirically for each primer set by varying $MgCl_2$ concentration (2 or 4 mM) and temperatures. In each case, the lowest $MgCl_2$ concentration and the highest temperature, at which bands could be seen, were chosen for sequencing. The temperature conditions (see Table 2.1, Chapter II) varied from the calculated Tm (See materials and methods) for each primer set as follows (+ is higher than the Tm, - is below, all °C): G1:+2, G2:+1, G3:-1, G4:-2, G5:+5, G6:+1. For set G1-G4, bands could be seen at 2 mM $MgCl_2$ whereas for G5 and G6, bands could only be seen at 4 mM $MgCl_2$. The resulting PCR products for primer sets G1-G4 (Fig. 5.1 d), set G5 (Fig. 5.1 e) and set G6 (Fig. 5.1 f) all contained bands only in the +RT lanes, whereas nothing was seen in the –RT or control lanes, indicating that all PCR products were due to cDNA and not contaminating gDNA or non-specific amplification. All observed bands were within the expected range, however primer set G3 and G4 detected two differently sized transcript populations. These were the only ones, where it had been necessary to use an annealing temperature below the calculated Tm of the primers, as higher annealing temperatures did not produce any bands.

5.2.4 Sequence analysis of clones

It should be noted that the strain of *P.yoelii* (17X A) used in our experiments differed from the sequenced (1.1) strain (Carlton et al., 2002) and thus some genuine sequence differences were observed. Overall, this was not found in the majority of genes, but it was taken into consideration when a stretch of mismatches occurred in between two highly scoring match regions and the sequence quality was good in the entire region. In all cases, these differences were so small that it did not interfere with identifying a particular gene by BLASTN, although it did lower the BLASTN output parameters in these few cases.

A total of 96 clones were obtained from RT-PCR with the G1-G6 primers; G1/G1 hot: 32, G2:15, G3: 15, G4: 13, G5: 10, G6:11. These were sequenced and the sequence quality was assessed (see materials and methods). BLASTN was performed against the CDS database at www.tigr.org, and the two first hits were retrieved to evaluate if there was more than one perfect match in the database, which could lead to ambiguity in the identification of genes. In total, the following number of clones obtained with each primer set were analysed further: G1: 18, G1-hot: 14, G2: 15, G3: 14, G4: 13, G5: 10, G6: 10.

By BLASTN against the *P.yoelii* annotated gene database at: www.tigr.org with the sequences from the 96 clones, the average length of match was 426 bp, and the average percentage match was 89%. This made it possible to assign all these 96 sequences to annotated *yir* genes. Some *yir* genes had almost identical copies within the genome, suggesting a recent duplication event. In this analysis, 20 clones were found to have only small differences in their E values when comparing the best and second best BLASTN hits. Not surprisingly, for 19 out of these 20 clones, the second best hits belonged to the same group/supergroup as the best hit. In one case a SG5 top hit was most similar to a SG1 second hit, and this probably relates to the finding that the SG1 BLASTN analysis had most of its intermediate homologues in SG5 (see Chapter IV). This sequence was classified according to its top hit (SG5).

5.2.5 Evaluation of primer performance

The performance of the primers were evaluated by investigating two parameters, namely:

- Primer bias, defined as the number of times the same sequences occurred in clones from a PCR reaction with the same primer (Fig.5.2 a).
- Supergroup specificity, defined as the percentages of clones belonging to the expected supergroup (Fig.5.2 b).

Bias normally means that there is a preference towards detecting a specific sequence more than others in a set of clones. This is generally caused by preferential amplification of some sequences in the PCR due to either preferential amplification of better matching sequences or differences in transcript levels. Preferential ligation of sequences with smaller sizes into the sequencing vector can also introduce bias. Here, this parameter was measured as the percentage of different sequences out of the total number of clones analysed for each primer set: (no. of clones containing gene X/total number of clones) x100%. Specificity is normally considered to be the ability of a primer to selectively amplify a high matching sequence over other lower matching sequences. Here, specificity was measured as the percentage of clones containing transcripts from genes within the expected supergroups (see Table 5.2). A high level of bias is indicative of selective amplification of a few sequences, whereas the successful amplification of sequences from the expected repertoire indicates specificity. A significant level of bias in the PCR reactions with primer G4 and G6 (Fig. 5.2 a), as a high number of repeated sequences were seen (around 60 and 42% respectively). Primer G5 had no detectable bias as all clones contained different sequences. Use of hot-start PCR with primer set G1 (G1 hot) led to a higher level of bias, as more clones (around 35%) contained the same sequences compared to around only 19% of

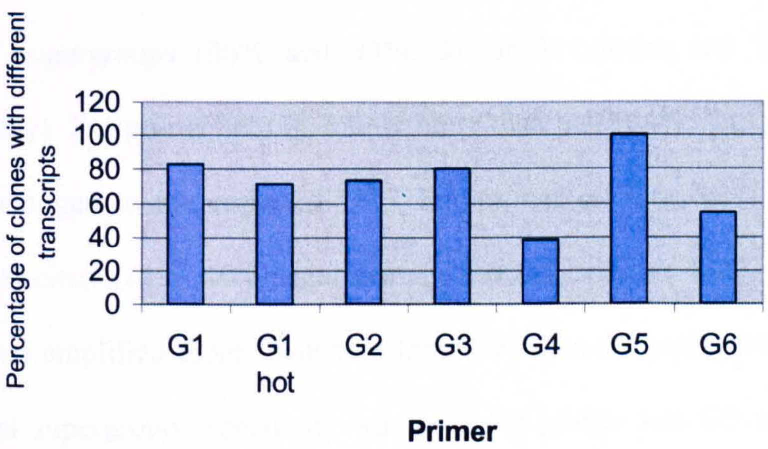
Figure 5.2

Evaluation of primer performances

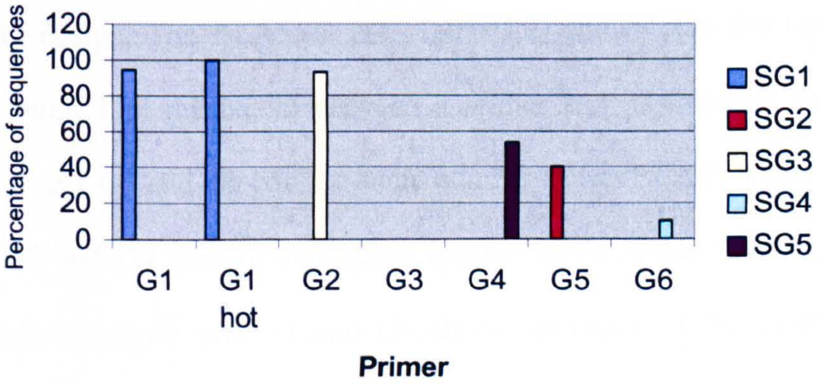
a) Primer bias. Percentage of different transcripts of the number of clones (number of transcripts/sequenced number of clones x100) obtained from each of the RT-PCR reactions. A low percentage indicates that the same sequences were identified in a high number of clones and a high percentage that the number of sequences and clones were more on par.

b) Primer specificity. Percentages of clones belonging to the expected supergroups, as estimated from BLASTN analysis. The primers were expected to be specific towards the following supergroups: G1: SG1, G2: SG3, G3:SG4, G4: SG5, G5: SG2 and G6: SG4. These expectations were based on high scoring BLASTN matches shown in Table 5.2.

a)


Percentage of clones with different transcripts
120
100
80
60
40
20
0
G1
G1 hot
G2
G3
G4
G5
G6
Primer

b)


Percentage of sequences
120
100
80
60
40
20
0
G1
G1 hot
G2
G3
G4
G5
G6
Primer
SG1
SG2
SG3
SG4
SG5

clones with a normal Taq DNA polymerase. Primer sets G2 and G3 had relatively low levels of bias.

The specificity of the primer sets were evaluated as the percentage of the total number of clones belonged to the expected supergroups (Fig 5.2 b). Primer set Gl and G2 had performed very well in specifically amplifying transcripts from their expected supergroups (96% and 93%, 31 of 31 clones and 14 of 15 clones respectively). For primer set G4, a little more than half (54%, 7 of 13 clones) of the clones belonged to the expected SG5. For primer set G5, only 40% of the total number of clones (4 of 10 clones) belonged to the expected SG2. However, primer set G5 had amplified more sequences from the expected supergroup than any other individual supergroup. Specificity was poor for primer sets G3 and G6, as a low proportion of clones from the expected SG4 were observed (0% and 10%, 0 of 14 clones and 1 of 10 clones respectively). The low success rates for in particular primer sets G3 and G6 could be caused by several factors:

1) Primer design. The BLASTN list (Table 5.2) shows only the top matches (maximum 1 nt mismatch) between a primer and its intended supergroup. Primer sets G3 and G6 had the same number of top BLASTN hits as primer sets G2 and G4, which were more specific in detecting their supergroups. Therefore, primer sets G3 and G6 should theoretically be as efficient as primer sets G2 and G4.

2) PCR/cloning bias due to size differences. As seen in Chapter IV, genes in SG4 and SG5 were found to be generally larger than genes in the other supergroups. Bias could have occurred due to selective amplification or

preferential cloning of smaller sequences outside SG4, which was the supergroup where two primers (sets G3 and G6) failed to perform as expected. Comparatively, the average sizes of exon 2 in the detected genes were: SG1: 807, SG2: 747, SG3: 786, SG4: 855 and SG5: 860 (all in bp). As the larger gene sizes for SG4 and SG5 were earlier found (see Chapter IV) to be caused by extra sequence just before the end of the second exon, this would have been included in the amplified regions with these primers. This could have led to bias at the level of amplification or cloning. This could also have affected primer set G4 that were designed towards SG5, and where only 54% of the clones were found to belong to this supergroup. Indeed, two differently sized populations of transcripts were seen on the gel (Fig. 5.1 d) with primer sets G3 and G4. From this, similar intensity bands were seen in the two populations with primer set G3, whereas the high molecular band was more intense than the small molecular band for primer set G4. This could also explain why more SG5 transcripts were detected with primer set G4 than SG4 transcripts with primer set G3. However, the inclusion of primer set G6 to detect SG4 transcripts did not lead to detection of these.

3) Significant variation in the transcript level. It cannot be excluded that the transcripts present unevenly represented the supergroups. Especially for SG4, where two primers consistently failed to amplify sequences, it is possible that these primers had amplified sequences from outside their expected repertoire because there were not many available targets in the sample.

4) PCR conditions. If the PCR conditions had been sub optimal in terms of stringency, the primers could have amplified poorer matching sequences

outside their expected repertoire. This could have been the case for primer sets G3 and G4, which were found only to produce bands at annealing temperatures below their calculated Tms. However, in terms of $MgCl_2$ concentration, the conditions were similar to the highly successful primer sets G1 and G2. The situation was opposite for primer set G6, which was against SG4, as was set G4; for this set, the annealing temperature was above the expected Tm, but it was only possible to obtain bands at 2 mM $MgCl_2$.

In total, the following numbers of different transcripts (when removing the number of repeated transcripts in several clones) were detected with each primer set: G1:15, G1-hot:10, G2:11, G3:12, G4:5, G5:10, G6:6. Some of the sequences amplified with sets G4 and G6 were repeated in several clones, demonstrating that these two primers had amplified a more restricted sequence repertoire than the other primers. This could have been caused by any one, or a combination, of the factors mentioned above. Ideally, all the primers should have been tested on genomic DNA prior to this analysis, as this would have enabled a true evaluation of primer performances. Unfortunately, this could not be carried out since a splice site spanning primer was used precisely to avoid any amplification of genomic DNA. Since high quality RNA was needed for this analysis, DNase treatment was not an option. Therefore, using the splice site spanning primer was the only viable option available. In addition, the exon 2/exon 3 splice-site primer was within the most conserved part of the gene, even more compared with the third exon which consisted of two slightly different types. If exon 3 primers had been used, they would allow for detection of spliced transcripts, but at least two types had to be used together with each exon 2 primer. In addition, the risk of competition between gDNA and cDNA amplification in any RT-PCR approach would compromise the value of data generated with this primer.

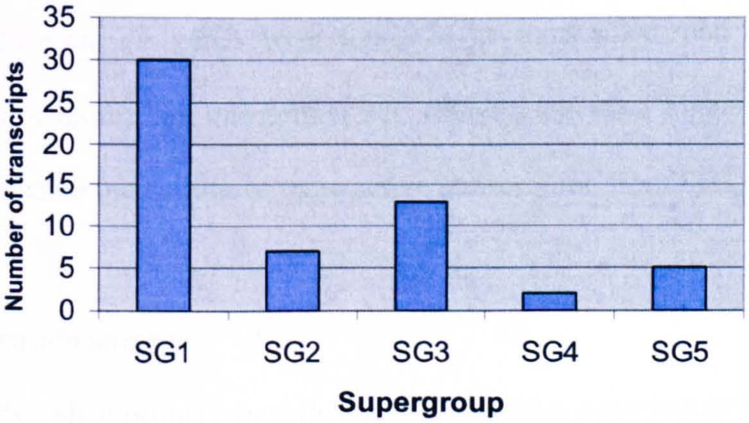
5.2.6 Analysis of transcribed genes.

While not all the primers performed as expected, they still enabled us to detect transcripts from all supergroups and show that a large diversity existed for the transcribed *yir* genes in the sample as the supergroups represented the most divergent *yir* genes. To determine how many transcripts were detected in total, all sequences found in multiple clones from the same RT-PCR reaction were designated as only one transcript. Overall, 57 different transcripts were detected from the blood stage sample and in some cases more then one clone contained the same transcript (See Fig. 5.2 a). However, 8 were detected by RT-PCRs using different primers, with one transcript detected by 4 primers, another by 3 primers and the others each by two different primers. The 57 transcripts were placed into their respective supergroups, however seven transcripts could not be easily placed. Supergroup location was determined using BLASTN analysis. For each of the 7 transcripts that could not immediately be located to a supergroup, the top three hits on the BLASTN analysis consistently belonged to the same supergroup, this allowed the grouping of all transcripts accordingly. Most of the transcripts detected in this analysis belonged to supergroup 1 (Fig. 5.3), however as this is also the largest of the supergroups, this may not be surprising. In total, 10% of SG1 genes were detected, whereas 15% of SG3 genes were detected. Of the 13 transcripts detected in SG3, 11 were detected with primer set G2, whereas 2 were detected with primer set G3. It was important to establish whether the transcripts detected by multiple primers reflected a higher transcription level. The eight transcripts that were detected in more than one RT-PCR reaction were numbered 1 to 8 and belonged to the following supergroups: PY03944 (SG1), PY07538 (SG1), PY05026 (SG1), PY05681 (SG1), PY06732 (SG4), PY07466 (SG2), PY05923 (SG2), PY02564 (SG1).

Figure 5.3

Transcription from supergroups

The number of transcripts detected in each of the supergroups is shown.


35
30
25
20
15
10
5
0
Number of transcripts
SG1
SG2
SG3
SG4
SG5
Supergroup

It was investigated which primer detected each of these transcripts and also if each particular transcript originated from a gene within the BLASTN repertoire of that particular primer (Fig. 5.4). From this it was clear that in some cases, the transcripts were detected with primers that theoretically should not detect them. This along with the fact that these transcripts were detected in multiple RT-PCRs would indicate that each of these 8 transcripts may represent an especially highly transcribed *yir*.

5.3 Summary and discussion

From the analysis in this chapter, it is clear that many transcripts are present in the blood stages of *P.yoelii*, and that transcripts from all the supergroups are present. In total, 57 different *yir* genes were found to be transcribed, and 8 of these were detected with more than one primer set, which suggests a higher transcript level. There were some indications of more active transcription from SG1 and SG3.

Immune evasion strategy

In this study, when primers were designed for specific supergroups it led to detection of active transcription from all supergroups. This indicates that a large number of *yir* genes are actively transcribed in the blood stages. In *P.falciparum*, the *var*-encoded Pfemp proteins switch to other adhesive variants at around 2.4% per generation (Roberts et al., 1992), and different Pfemp proteins exhibit very little antibody cross-reactivity (Newbold et al., 1992 and Bull et al., 1999). Although multiple *var* genes are transcribed and translated, dominant full length *var* transcript are present during an infection (Taylor et al., 2000 and Noviyanti et al., 2001). *Var* genes have been proposed to have distinct on and off rates (Horrocks et al., 2004). Why one *var* gene becomes dominant is probably caused by differences in these rates during the initial

Figure 5.4

Dominant transcripts and primer repertoire

In the left column are shown the eight transcripts detected more than once (numbered 1 to 8). In the top row, the primers are indicated. G1.1 and G2.2 is the same primer, but PCR was performed without (G1.1) or with (G1.2) a hot start enzyme. Detected transcripts are indicated by boxes, and 🟦 indicate that the transcript is detected, and is within the repertoire of the used primer, whereas 🟥 indicate that the transcript is detected but is not within the repertoire of the used primer.

a)

b)

	G1.1	G1.2	G2	G3	G4	G5	G6
1	🟦	🟦		🟦			🟥
2	🟦	🟦		🟦			
3	🟦	🟦					
4	🟦			🟦			
5						🟥	🟥
6				🟥	🟦		
7					🟥	🟦	
8	🟦	🟦					

a) Primer

b) Transcript number

stages of an infection, when the number of parasites is relatively small. However, the low level switching gives *P.falciparum* the ability to maintain the infection if the dominant type is targeted by the host's immune system. The involvement of the *var* intron as an alternative promoter that could generate sterile, truncated *var* transcripts (Calderwood et al., 2003) probably explains why all but one of the ring stage transcripts are truncated (Taylor et al., 2000). Overall, the *var*/Pfemp model suggests an immune strategy based on semi-sequential expression where, if the dominant type is removed by the host's immune system, other variants would allow the infection to persist.

Extrapolating from the rather limited study performed in this thesis, given the size of the *yir* repertoire, transcription of several hundred *yir* genes would not be surprising. It is of course an assumption that all transcribed *yir* genes are also translated, since a mechanism exists for generating truncated *var* genes (as described above). The primers used in this study were located in exon 2/exon 3 and would thus not detect 5′ truncations. The 5′RACE analysis, (Chapter III), only detected full-length *yir* genes in the blood stages. However, (as will be described later) a mechanism does exist for *yir* truncations, but at a very low level of occurrence. Therefore, it is assumed that the majority of the transcripts detected in this analysis are translated. These data were part of a study that investigated the effect on *yir* transcription upon reinfection in immunocompetent and immune-compromised mice (Cunningham ,2005). The study showed that YIR proteins are expressed at the iRBC surface in the schizont stages, and that the transcription profile of a subset of SG1 genes changed consistently (but not completely) upon reinfection of immunocompetent mice, whereas no changes occurred in immuno-compromised mice (Cunningham et al., 2005). This emphasize that some YIR proteins are targeted by the host's immune system, and this is probably because they are expressed at a relatively higher level. In this thesis, 8 *yir* transcripts were found in more than one RT-PCR reaction using different primers, and are proposed to be part of a set of highly expressed YIR proteins. Therefore there might be an initial set of YIR proteins occurring by default, but accompanied by a vast repertoire of lower expressed variants. However, the frequency for these proposed dominant transcripts only accounts for 14% (8 out of 57) of the detected repertoire. Although this resembles the *var* genes at first glance, this would be a fundamentally different immune evasion strategy, as it appears as if *P.yoelii* does not economize with its repertoire in a manner similar to *P.falciparum*.

One important question, which was not addressed in this study, was the speed with which a large repertoire of *yir* genes becomes established during an infection. This depends on how many *yir* genes are transcribed per individual parasite, and how many different *yir* genes are transcribed per cycle. As the present study used an inoculum of a very high number of parasites, a large repertoire would already have been established from the onset of infection. Single cell infection studies are needed to determine just how quickly a large *yir* repertoire is established.

Transcription of different supergroups

It is thought that transcripts from SG4 and SG5 could be present at lower levels, and they are therefore only detected using less stringent PCR conditions. This is not a reflection of the total number of genes in SG4 and SG5 compared to the other supergroups, since there are a comparable number of genes in SG3, where many transcripts were detected when primer set G2 was used. Unfortunately, at these less stringent conditions, the primers (G3 and G6 in particular) fail to successfully differentiate between transcripts in the expected supergroups and other supergroups.

Intuitively, the probability for a primer to amplify a sequence is mainly dependent on two conditions: the chance of encountering a sequence and the strength of the match. If a strong match occurs very rarely at stringent PCR conditions, due to low transcript levels, little amplification will occur. For both G3 and G6, the failure of applying very stringent conditions was observed. This could therefore be due to fewer available transcripts from these supergroups.

Cloning bias could also explain why relatively few transcripts were seen in SG4 and SG5. Both primer Set G3 and G4 products (Fig. 5.1 d) consisted of two transcript populations as seen on the gel. For primer set G4, a more intense larger band was seen, and indeed this primer performed better than primer set G3 in identifying transcripts from its expected supergroup. However, it was only realized at the end of this PhD thesis that size differences did exist in the supergroups, and therefore no attempts to cut out bands were performed during this experiment.

Studies of the *var* genes have found that subtelomerically located *var* genes interacted with the Sir2 protein and were located in telomeric clusters when inactive (Ralph et al., 2005). This contrasted to central *var* genes, which did not interact with the Sir2 protein nor did they associate with the telomeric cluster (Ralph et al., 2005). Central *var* genes are not constitutively active, but appear to be regulated by a Sir2 independent mechanism (Ralph et al., 2005). In addition, *var* genes with different 5′ upstream sequences (Ups A/B and C-type) interacted differentially with nuclear proteins in a distinct manner (Voss et al., 2003). Therefore, *var* genes are regulated by different mechanisms depending on the chromosomal localization.

As the *yir* repertoire also consist of different supergroups with distinct distributions on annotated telomeric and subtelomeric contigs, this could indicate differential chromosomal localization of these groups (see Chapter IV). If this is the case, they could be regulated in a similar manner as the *var* genes. In this study, more transcripts were detected from SG1, SG2 and SG3 than the two highly subtelomerically located SG4 and SG5. Even though cloning bias could have played a role in skewing the detected repertoire, it is thought (as discussed above) that there are differences in the transcriptional activity of the different supergroups. If so, this could indicate that different epigenetic mechanisms operate on the different supergroups.

There are, however several limitations to this analysis; generally RT-PCR is not very suitable for quantification as it is biased in a number of ways. In this study, the primers especially for SG4 and SG5 failed to amplify a diverse number of sequences from these supergroups. This can, at least partially, be explained by cloning bias for SG5, but for SG4, it is more likely that less transcription occurs from this supergroup, which was why only less stringent PCR conditions worked with both primer sets (G3 and G6). Therefore, it is an open question whether there is a link between the localization of the supergroups and their transcriptional activity in the blood stages. The present analysis, although several valid objections can be raised against this, suggests that there might be a preferential transcription from some supergroups.

Chapter VI

Single cell RT-PCR

6.1 Introduction

A number of studies have investigated transcription of *Plasmodium* multigene families in single infected erythrocytes. This has given valuable information of how many transcripts are present per individual cell and indicated on the maximum number of proteins that could be expressed per cell, as well as highlighting the degree of transcriptional restriction. For the *var* genes, up to 15 transcripts were detected in ring stage parasites, whereas only one was detected in the trophozoite stage (Fernandez et al., 2002 and Scherf et al., 1998 and Chen et al., 1998). A maximum of three *stevor* transcripts could be detected in late schizont stage parasites (Kaviratne et al., 2002). Studies of another multigene family, the *Py235* genes in *P.yoelii* showed that whereas most trophozoite stage parasites only transcribed one gene, several genes were transcribed in the schizont stage (Preiser et al., 1999). In addition, each merozoite within the schizont transcribed only one gene, which is a prime example of how clonal expression can occur (Preiser et al., 1999). All these studies emphasises, that despite the presence of a large repertoire of genes in the studied multigene families, strong transcriptional regulatory mechanisms ensure that only a limited subset of these genes are active in a parasite within the iRBC. In two of the above-mentioned studies, it was also found that the degree of transcriptional repression depended on the life cycle stage of the parasite.

6.1.1 Objectives

Several standard RT-PCR approaches were found to be unsuccessful in detecting *yir* transcripts from single infected erythrocytes (6.2.1). However, two commercially available kits allowed for detection of transcripts from single infected erythrocytes at the schizont stage (6.2.2 to 6.2.3). Trophozoites were also investigated with one of these kits (6.2.4); however, an unexpected problem was encountered when using this kit.

6.2 Results

6.2.1 Set-up of single cell RT-PCR

Initial studies on serially diluted RNA and cells showed that reverse transcription with random hexamers or gene specific primers, followed by PCR with splice site spanning primers could not detect transcripts from less than 600 infected cells. Therefore it was decided to amplify the RNA prior to the PCR step. Two methods for RNA amplification were tried, namely the aRNA Message Amp kit from Ambion and the SuperSMART cDNA synthesis kit from BD (formerly Clontech).

6.2.2 Single cell RT-PCR using the Message Amp™ kit from Ambion

In short, the Message Amp kit technology allows for linear amplification of double stranded cDNA using the T7 promoter, which is incorporated into the cDNA during reverse transcription (see materials and methods). The resulting anti sense RNA (aRNA) can then be DNase digested to remove residual cDNA and gDNA and column- purified. The manufacturer claims that up to a 1000 fold increase in the level of RNA (aRNA) can be achieved. For this application, the resultant aRNA was reverse-transcribed using random hexamers and SuperScript II reverse transcriptase (see materials and methods). Here (Fig. 6.1 a and b), a nested approach using an outer primer set in exon 2 and exon 3 followed by an internal primer set in exon 2 and the exon 2/exon 3 splice site was used. A positive signal was obtained from one cell (Fig. 6.1 b C1). In addition to the nested PCR, a semi-nested PCR approach was used in which the same exon 3 primer was used (Fig. 6.1 b, set D). This was done to investigate if the sequence was spliced between exon 2 and exon 3. This approach enabled us to test by sequencing if the sequence was spliced, which would be a genuine characteristic for cDNA as opposed to gDNA. Sequencing of 12 clones all revealed a correctly spliced sequence between the second and third exon. All the sequences were highly similar (>90%) to a *yir* gene on contig 1543. This showed that a genuine transcript was detected from this single micromanipulated schizont.

Despite this positive result, it was not possible to reproduce it with eight other cells. Because of this low success ratio, a method relying on an exponential rather than a linear amplification of RNA was chosen for further analysis, as this might lead to more positive cells being detected.

6.2.3 Single cell RT-PCR using the SuperSMART™ kit from BD

SMART technology (see materials and methods) relies on the incorporation of a primer sequence tag during the first strand synthesis with a long oligo dT. Template switching occurs when the reverse transcriptase reaches the end of the transcript, whereby another oligo, the SMART oligo, is incorporated into the cDNA. The resulting cDNA then has identical primer sequences at the 5′ and 3′ ends, which are then used for long distance PCR amplification (see materials and methods). The manufacturer claims that as little as 2 ng total RNA can be used as starting material.

In an initial test the cDNA synthesis and amplification was performed as recommended by the manufacturer using one whole single iRBC at the schizont stage (called C2). It was tested in a semi-nested PCR, if spliced transcripts would be detected (Fig. 6.1 c). A strong band was observed from this cell (Fig. 6.1 c, set D).

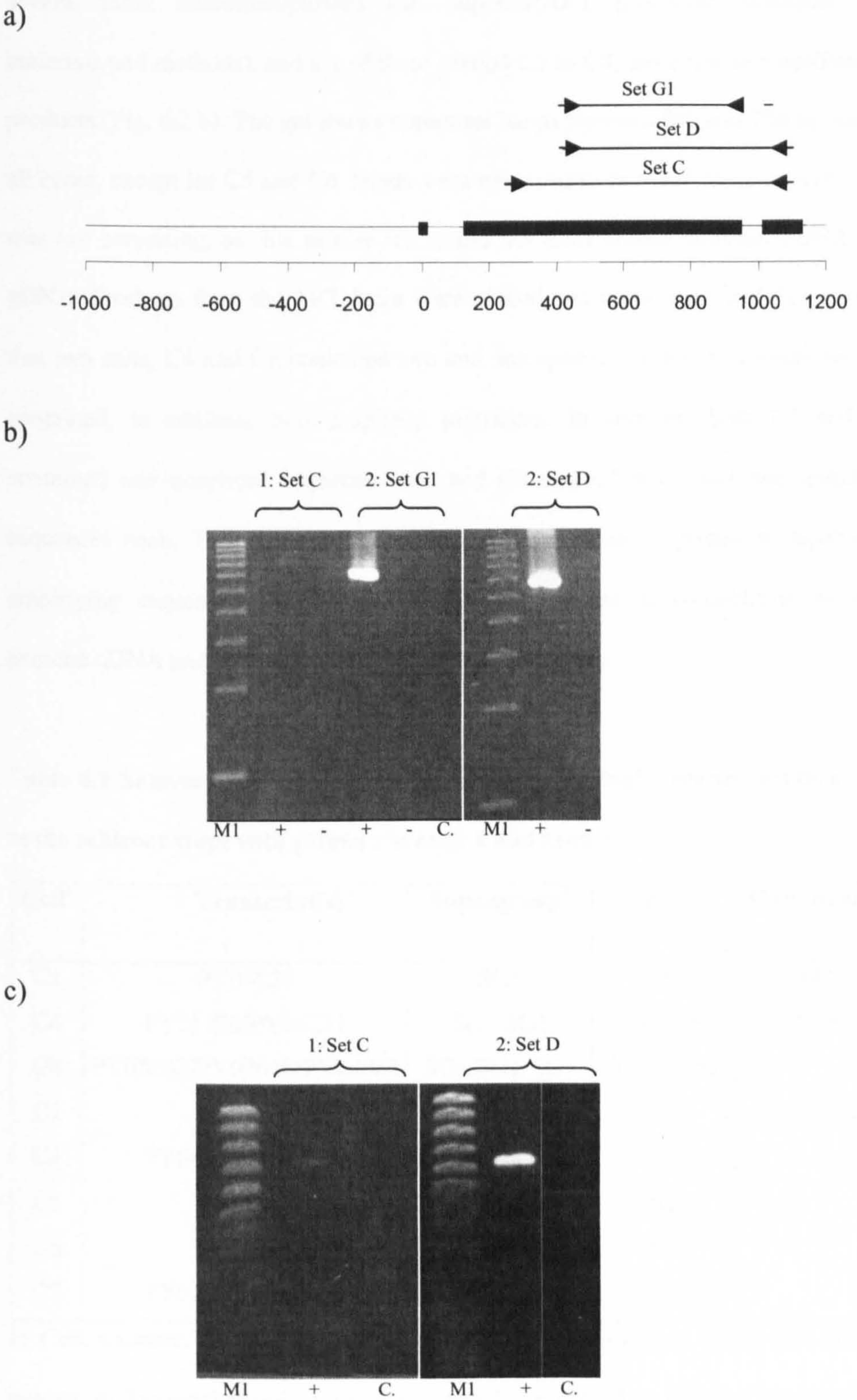
Figure 6.1

Initial tests of the two kits

a) Location of the used primer pairs. The expected size ranges of products amplified with the different primer sets were: Set C: 660-760 bp (cDNA), Set D: 510-610 bp (cDNA), Set G1: 500-600 bp.

b) Amplification of aRNA using the MessageAmp kit and reverse transcription was performed on a single infected erythrocyte at the schizont stage separated into +RT and –RT samples (see materials and methods for micromanipulation and the MessageAmp kit). PCR was performed with primer Set C. A 1:100 dilution of this was then used in a nested (Set G1) and semi-nested (Set D) PCR reaction. Intense ~600 bp bands were seen after the nested and semi-nested PCRs when comparing to the 100 bp marker (M1). Nothing was seen in the –RT or control lanes. The product from the semi-nested PCR was cloned and sequenced.

c) Amplification of cDNA using the SuperSMART kit was performed on a single infected erythrocyte at the schizont stage (see materials and methods for micromanipulation and the SuperSMART kit). This cell was not separated into a + and –RT sample prior to cDNA amplification. PCR was performed with primer Set C. A 1:100 dilution of this was then used in a semi-nested (Set D) PCR reaction. An intense band of ~600 bp was seen after the semi-nested PCR when comparing to the 100 bp marker (M1). Nothing was seen in the control lanes. The product from the semi-nested PCR was cloned and sequenced.


a)
Set G1
Set D
Set C
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
1: Set C
2: Set G1
2: Set D
M1
+
-
+
-
C.
M1
+
-
c)
1: Set C
2: Set D
M1
+
C.
M1
+
C.

Sequencing revealed it was spliced. The exon 3 primer was used in PCR reactions for several more micromanipulated and SuperSMART processed schizonts (see materials and methods), and six of these (called C3 to C8) gave rise to amplification products (Fig. 6.2 b). The gel shows numerous bands between 600 and 750 bp, and in all cases, except for C5 and C6, bands were observed in the –RT lanes as well. This was not surprising, as this primer set would not discriminate between cDNA and gDNA. Products from the +RT lanes were cloned and sequenced, and this showed that two cells, C4 and C8 contained two and one spliced transcript respectively. C8 contained, in addition, two unspliced sequences. In contrast, both C5 and C6 contained one unspliced sequence each and C3 and C7 contained two unspliced sequences each. This showed that while the exon 2/exon 3 primer is capable of amplifying sequences arising from mRNA, it also led to competition between genuine cDNA and gDNA sequences during the PCR step.

Table 6.1 Summary of sequenced PCR products in single infected erythrocytes at the schizont stage with primers in exon 2 and exon 3.

Cell[1]	Transcript(s)[2]	Supergroup[3]	Spliced[4]	Comments[5]
C1	PY04939	SG1	yes	1st aRNA
C4	PY01720/PY06131	SG1/SG1	Yes/yes	3 PCRs
C8	PY05242/PY07618/PY04082	SG1/SG1/SG1	Yes/no/no	
C2	PY04580	SG1	Yes	1st SMART
C3	PY06887/PY03089	SG1/SG1	No/no	
C5	PY04966	SG1	No	
C6	PY06907	SG1	No	
C7	PY02727/PY04419	SG1/SG1	No/no	

1: Cell number. 2: Detected transcripts. 3: Supergroup, to which transcripts belong. 4: Transcript spliced between exon 2 and exon 3 or not. 5: Comments.

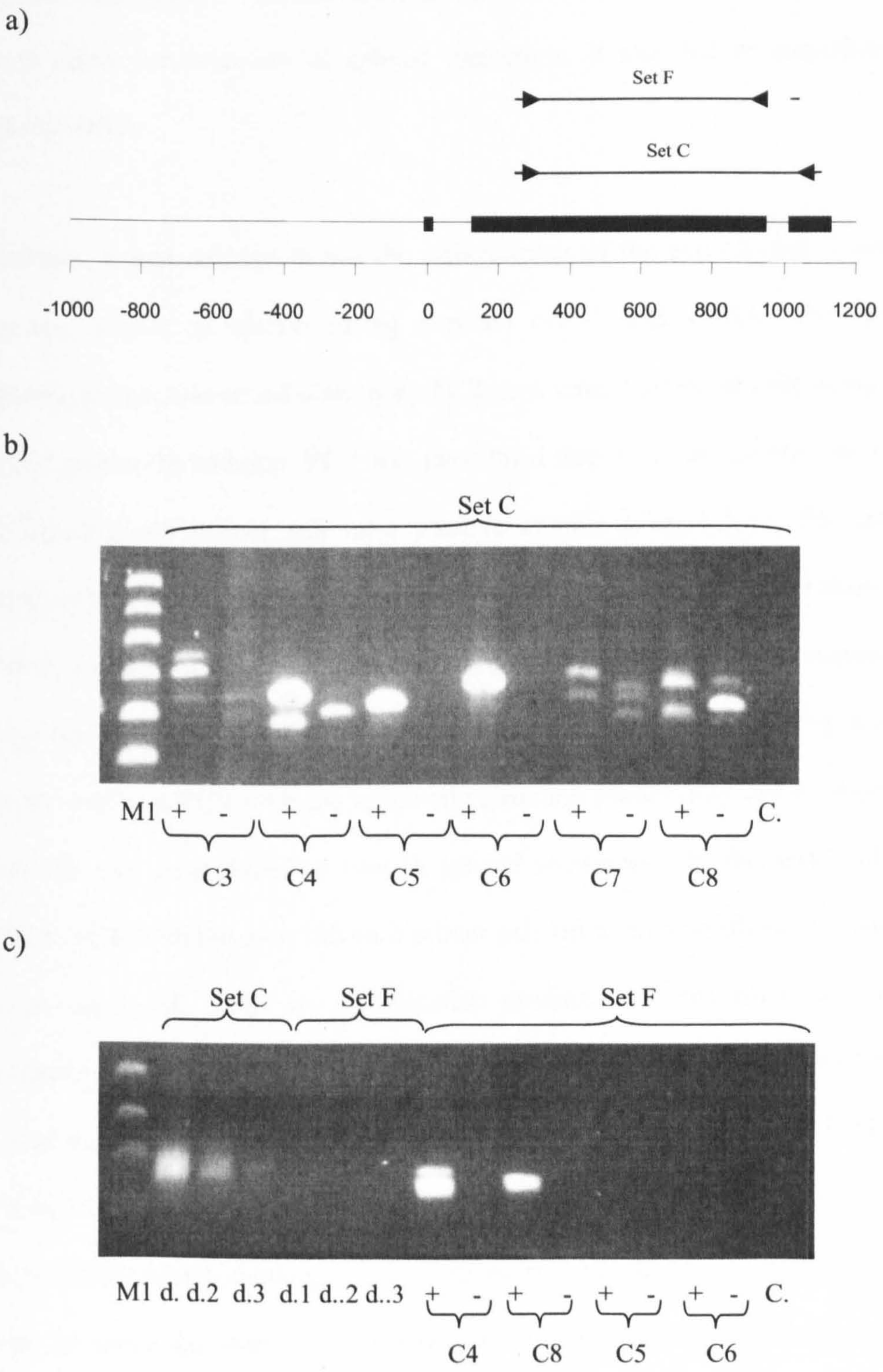
Figure 6.2

RT-PCR on single cells

a) Location of the used primer pairs. The expected size ranges of products amplified with the different primer sets were: Set C: 750-850 (gDNA) and 660-760 (cDNA) bp, Set F: 650-750 bp.

b) PCR was performed with primer set C on cDNA amplified with the SuperSMART kit from six single infected cells at the schizont stage (C3 to C8 and separated into + and – RT samples) picked by micromanipulation (see materials and methods). Bands ranging from 600 bp to 750 bp were seen in both + and – RT lanes when comparing to the 100 bp marker (M1). Nothing was seen in the control (C.) lane.

c) PCR with primer sets C and F was performed on genomic DNA from a dilution of cells (d.1 100 cells, d.2 10 cells and d.3 1 cell). Bands of around 780 bp (compared to 100 bp marker, M1) were only seen in the lanes containing products obtained from PCR with primer set C. As a test, PCR with primer set F was also performed on C4, C8 and C5, C6 (see above), which were found to contain spliced sequences (C4 and C8) and unspliced sequences (C5 and C6) respectively. Here, products were only seen from C4 and C8. Nothing was seen in the –RT or control (C.) lanes.


a)
Set F
Set C
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
Set C
M1 + - + - + - + - + - + - C.
C3 C4 C5 C6 C7 C8
c)
Set C
Set F
Set F
M1 d. d.2 d.3 d.1 d..2 d..3 + - + - + - + - C.
C4 C8 C5 C6

As is summarized in Table 6.1, four cells were found to contain spliced sequences and another four cells contained unspliced sequences. The spliced sequences were considered genuine transcripts, whereas the unspliced sequences were most likely to originate from gDNA. This showed that while the exon 2/exon 3 primer combination would allow for detection of spliced transcripts, it also led to amplification of genomic DNA.

Therefore, it was decided to test the effectiveness of the exon 2/exon 3 splice-site spanning primer in discriminating between cDNA and gDNA. The splice-site spanning primer was tested directly by PCR on a serial dilution of cells along with an exon 2 primer. In addition, PCR was performed directly of the gDNA dilution with the exon2/exon3 primer pair as a positive control. (Fig. 6.2 c). The splice-site spanning primer set was then tested for its ability to amplify a product from the two cells (C4 and C8 in Fig. 6.2 c) previously shown to contain spliced sequences and two other cells that did not contain spliced sequences (C5 and C6 in Fig. 6.2 c). The test showed that PCR with the splice-site spanning primer only led to amplification from the two cells found to contain spliced sequences. On the serial dilution of gDNA, PCR with the exon2/exon 3 primer pair led to an amplification product from as low as 1 cell, while no amplification product was seen when the splice site spanning primer was used. Therefore, when using rather stringent conditions this primer combination was sufficient to discriminate between gDNA and cDNA. The cDNA from the four schizonts (one generated with the aRNA method and three with the SuperSMART technology) that contained spliced transcripts, were then amplified with the splice site spanning primer (Fig. 6.3 b). While a weak band can be detected (Fig. 6.3 b) from the aRNA amplified schizont (C1), two strong bands from one SuperSMART amplified schizont (C4), and one strong band from each of the

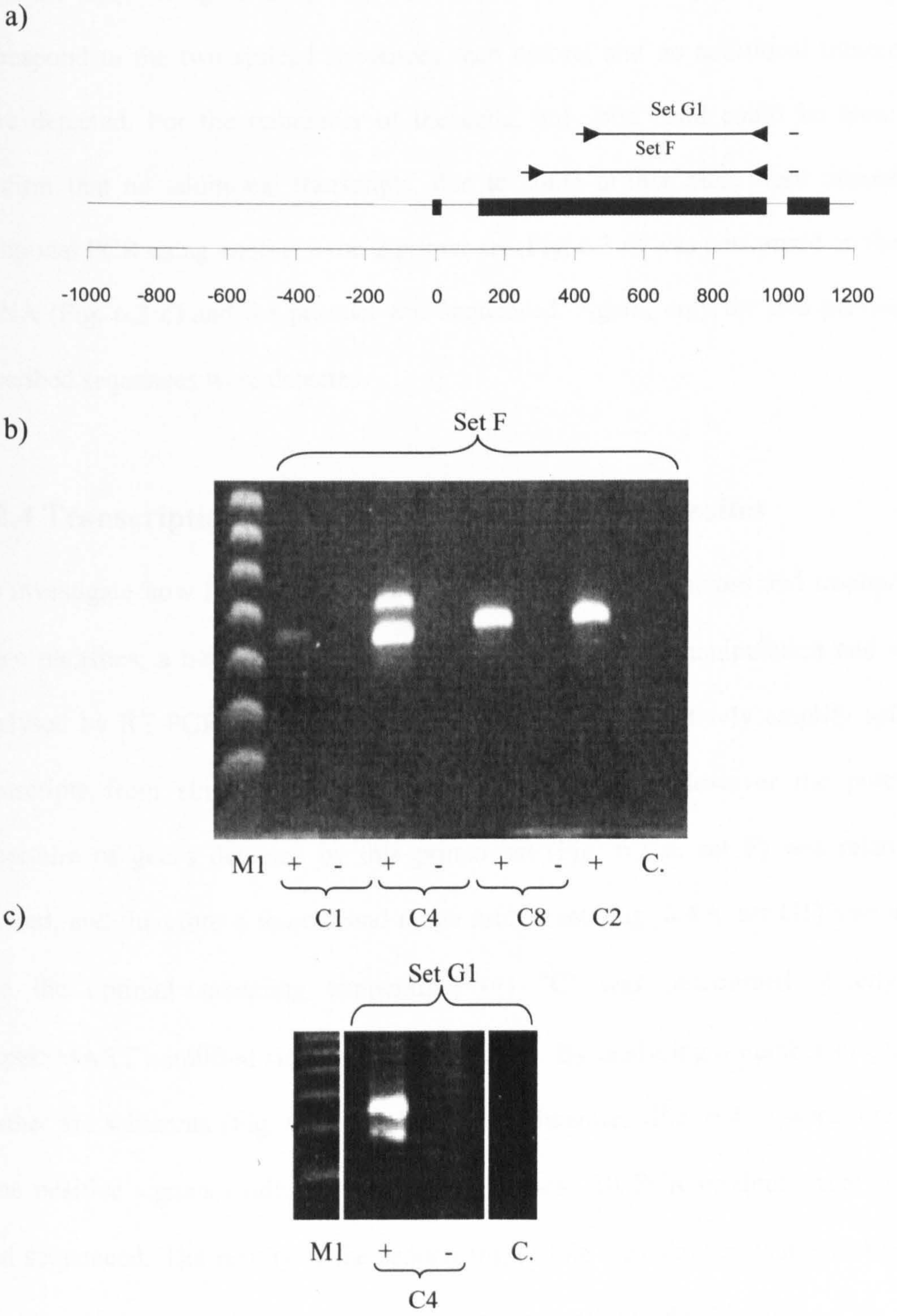
Figure 6.3

RT-PCR on four cells (C1, C2, C4 and C8) containing spliced sequences

a) Location of the used primer pairs. The expected size ranges of products amplified with the different primer sets were: Set F: 650-750 bp and G1: 500-600 bp.

b) PCR was performed with primer set F on four single infected schizont stage cells (C1, C4, C8 and C2 see above), which were found to contain spliced sequences. C2 was not separated into a + and – RT sample (see above). For C1, C8 and C2 single bands between 600 and 700 bp were seen (compared to the 100 bp marker, M1), whereas two bands of around 650 and 700 bp were seen for C4. Nothing was seen in the –RT or control (C.) lanes.

c) PCR with primer set G1 was performed on C4 and two bands of around 600 and 650 bp were seen (compared to the 100 bp marker, M1). This PCR product was cloned and sequenced. Nothing was seen in the –RT or control (C.) lanes.


a)
Set G1
Set F
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
Set F
M1 + - + - + - + C.
C1 C4 C8 C2
c)
Set G1
M1 + - C.
C4

additional two SuperSMART amplified schizonts (C2 and C8) were seen. No –RT was present for C2 as the cDNA from this sample was generated from a whole single Schizont. Sequencing of the product from C4 confirmed that the bands did indeed correspond to the two spliced sequences seen before, and no additional transcripts were detected. For the remainder of the cells, only one band could be seen. To confirm that no additional transcripts, due to some primer bias, were missed an additional PCR using another exon 2 primer set (Fig. 6.3 a) was performed on the C4 cDNA (Fig. 6.3 c) and the product was sequenced. Again, only the two previously described sequences were detected.

6.2.4 Transcription in single Schizonts and Trophozoites

To investigate how many transcripts could be detected in schizont and trophozoite stage parasites, a new batch of cells were picked by micromanipulation and were analysed by RT-PCR. The splice site primer is able to selectively amplify spliced transcripts from single cells under stringent conditions. However the potential repertoire of genes detected by this primer set (Fig. 6.3 a, set F) was relatively limited, and therefore a more broad range primer set (Fig. 6.4 a, set G1) was used, and the optimal annealing temperature (41 °C) was determined directly on SuperSMART amplified single cells (not shown). By analysing a number of cells, a further six schizonts (Fig. 6.4 b) and seven trophozoites (Fig. 6.4 c) were found to give positive signals exclusively in the +RT lanes. All PCR products were cloned and sequenced. The results of the sequencing (Table 6.2) showed that a number of *yir* transcripts were detected in more than one cell (Table 6.2, *italics*). In fact, sequences matching to PY02688 were found in four different cells, and sequences matching to PY04580, PY00351 and PY05828 where found in two different cells respectively.

Figure 6.4

RT-PCR on single Schizonts and Trophozoites

a) Location of the used primer pairs. The expected size range of product amplified with the used primer Set G1 was: 500-600 bp.

b) PCR with primer set G1 on six micromanipulated schizonts (Sz1 to Sz6) (see materials and methods) separated into +RT and –RT samples. Bands of around 600 bp were seen when comparing to the 100 bp marker (M1). Nothing was seen in –RT or control (C.) lanes.

c) PCR with primer set G1 on seven micromanipulated trophozoites (Tz1 to Tz7) (see materials and methods) separated into +RT and –RT samples. Bands of around 600 bp were seen when comparing to the 100 bp marker (M1). Nothing was seen in –RT or control (C.) lanes.

d) Alignment between a sequenced clone from Sz3 and the best matching contig and the corresponding annotated *yir* gene. The position of the splice site spanning primer (set G1 reverse) as it appeared in the sequence is boxed, and it can be seen from this that the primer had annealed to the third exon and the intron.

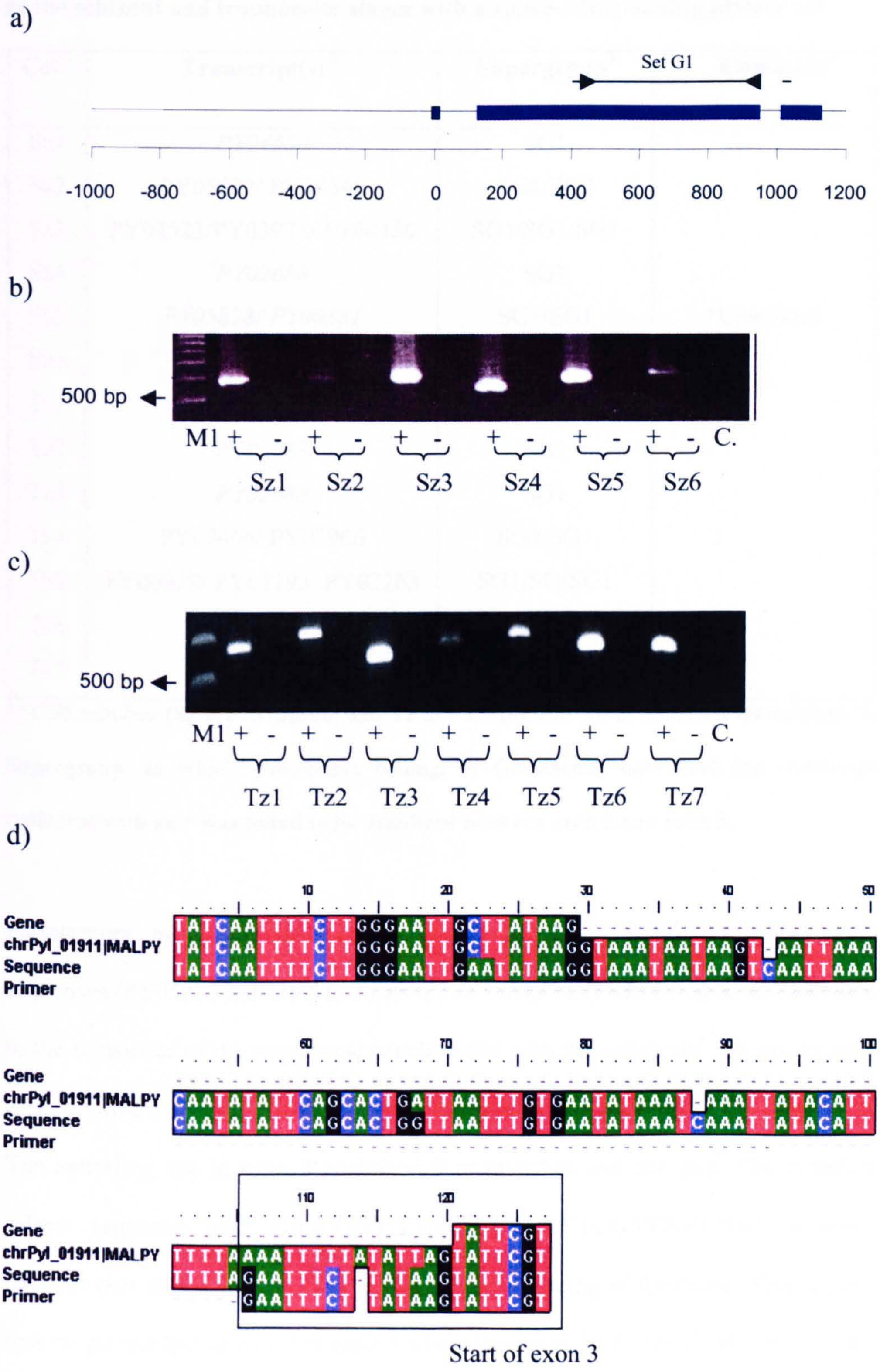

a)
Set G1
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
500 bp
M1
+
-
C.
Sz1
Sz2
Sz3
Sz4
Sz5
Sz6
c)
500 bp
M1
+
-
C.
Tz1
Tz2
Tz3
Tz4
Tz5
Tz6
Tz7
d)
10
20
30
40
50
Gene TATCAATTTTCTTGGGAATTGCTTATAAG
chrPyl_01911|MALPY TATCAATTTTCTTGGGAATTGCTTATAAGGTAAATAATAAGT-AATTAAA
Sequence TATCAATTTTCTTGGGAATTGAATATAAGGTAAATAATAAGTCAATTAAA
Primer
60
70
80
90
100
Gene
chrPyl_01911|MALPY CAATATATTCAGCACTGATTAATTTGTGAATATAAAT-AAATTATACATT
Sequence CAATATATTCAGCACTGGTTAATTTGTGAATATAAATCAAATTATACATT
Primer
110
120
Gene TATTCGT
chrPyl_01911|MALPY TTTTAAAATTTTTATATTAGTATTCGT
Sequence TTTTAGAATTTCT-TATAAGTATTCGT
Primer GAATTTCT TATAAGTATTCGT
Start of exon 3

Table 6.2 Summary of sequenced PCR products in single infected erythrocytes at the schizont and trophozoite stages with a splice-site spanning primer set.

Cell[1]	Transcript(s)[2]	Supergroup[3]	Comment[4]
Sz1	*PY04580*	SG1	
Sz2	PY05822/ PY06340	SG1/SG2	
Sz3	PY02523/PY03973/ *PY04580*	SG1/SG1/SG1	
Sz4	*PY02688*	SG1	
Sz5	*PY05828/ PY00351*	SG1/SG1	*Unspliced
Sz6	*PY05828*	SG1	
Tz1	*PY00351*	SG1	
Tz2	PY02298	SG1	
Tz3	*PY02688*	SG1	
Tz4	PY07466/ PY03966	SG2/SG1	
Tz5	PY06939/ PY03193/ PY02263	SG1/SG/SG1	
Tz6	*PY02688*	SG1	
Tz7	*PY02688*	SG1	

1: Cell number (Sz are Schizonts and Tz are Trophozoites). 2: Detected transcripts. 3: Supergroup, to which transcripts belong. 4: Comments, note that the transcript indicated with an * was found to be unspliced between exon 2 and exon 3.

Furthermore, by closely analysing the sequences, it was found that one of these sequences (PY05828 indicated in Table 6.2 as *unspliced) was not spliced, compared to the remainder of the sequences, which ended with the splice site. As can be seen (Fig. 6.4 d) for this unspliced sequence, the primer had annealed to the third exon. The annealing site in exon 3 contained 3 mismatches and one gap. The complete primer sequence was: **GAATTTCTTATAAGTATTCGT**TATTTGG, however only the part shown in bold were detected by sequencing of the clone. This suggest that the primer had annealed to exon 3 where there was 100% match of 7 nucleotides and another potential 100% match of 7 nucleotides in the third exon. In the intronic sequence, 9 out of 14 nucleotides matched with the primer. It is therefore likely that at the low stringency conditions used this primer allowed amplification of an unspliced sequence. However, this was not observed in any of the other analysed sequences, but as seen (Table 6.2), this gene was detected in two different schizont infected cells (Sz5 and Sz6 respectively).

6.3 Summary and discussion

A limited number of *yir* transcripts were detected in single *P.yoelii* infected RBC. A maximum of two transcripts were detected from a single schizont, and three other schizonts were found to contain only one transcript each. These transcripts were genuine as they were all spliced between the second and third exons in accordance with the gene model (see Fig. 3.1, Chapter III). In addition, one of these cells was also investigated with two other primer sets. Even when using the additional primers, the same two transcripts were consistently detected from this cell. This means that, within the repertoire of these primers, a maximum of two spliced transcripts were detected. In addition, up to two unspliced sequences were detected in a cell containing a spliced transcript. The origin of these unspliced is not clear, but if they were from unprocessed mRNA, this would show that a single cell could contain at least three transcribed genes.

In a recently published study of the *vir* gene transcription in *P. vivax*, single cells from mixed trophozoite and schizont populations were found to contain transcribed genes from at least two of the five VIR encoding subgroups through the use of subgroup specific primers (Fernandez et al., 2005). In the present study, 3 out of 4 schizonts only contained one transcript, and no more than two transcripts could be detected in the remaining schizont after performing PCR with a total of three different primer sets (Sets C, G1 and F). The reverse primers in these sets were specific for highly conserved parts of the gene (splice site and exon 3 respectively), and approximately 90-95% of all *yir* genes could potentially be amplified with these primers alone. The combined repertoire of forward exon 2 primers (forward set C+F and G1) covers around 60% of the *yir* repertoire, with redundant hits removed, as found by BLASTN analysis. However it was earlier seen (Chapter V) that less stringent PCR conditions allowed primer set G1 to amplify sequences from outside its repertoire. In this study, the PCR conditions for primer G1 were very relaxed (see materials and methods), and therefore more than 60% of the *yir* repertoire was likely to be detected. Also, as discussed earlier (Chapter V) relatively few transcripts were detected from SG2, SG4 and SG5 in a population of parasites using primers specifically designed to amplify transcripts in these supergroups. Therefore, it can be concluded with a probability of at least 60% that between 1 and up to 2 *yir* genes were transcribed per infected cell at the schizont stage in the analysed cells. However, it cannot be ruled out that the schizont containing two transcripts (C4 Table 6.1) originated from a multiple infection event, as it is not possible to distinguish between schizonts originating from a single or a multiple infection. In this study, all the transcripts detected in the different single cells belonged SG1, which was also the supergroup from which most transcripts were earlier detected (see Chapter V). Since SG1 contains 332 genes, there is a higher probability (40%) of picking cells transcribing genes from this supergroup than from any of the other supergroups. However, it could also be, as discussed in the last chapter, that some supergroups are more transcriptionally active than others.

From an immunological perspective, the following considerations were made. In this chapter, with at least 60% theoretical certainty, 1-2 transcripts were detected per single infected RBC. This means that a single parasite only expresses 0.23 % of the *yir* repertoire. In Chapter V, 57 transcripts were detected, and 8 of these (14%) were detected more than once by RT-PCR. These 14% dominant transcripts could be expressed and targeted by immune-mechanisms, while the 86% minor variants would have a lower probability for being recognized by the immune system (save for cross reactive antibodies). If such a scenario is the case, it becomes crucial for individual parasites to keep YIR expression at a low level. Mathematically, the probability for a single parasite not to express YIR proteins recognized by the immune system (y), would be: P (y)= 0.86^x, where x is the number of transcripts per cell. If the average cell in a population expresses 1 YIR protein, it would have an 86% chance of survival, if it expresses 5 YIR proteins, only 47% chance of survival, and if 10 YIR proteins are expressed, a mere 22% chance of survival. This example is just to emphasize, that even with a large multigene family, the importance gene silencing mechanisms are crucial for survival.

However, the finding of several identical sequences in totally different single cells was in stark contrast to the earlier finding (see chapter V) of a diverse transcribed repertoire from RT-PCR on total RNA. The single cells were obtained from different infections and contained several identical transcripts, whereas the total RNA was obtained from the same sample and only contained 8 repeatedly detected transcripts out of 57 (Chapter V). It seems quite unlikely that the single cells from different infections should exhibit less diversity in their transcribed repertoire than the transcribed repertoire in blood from a single infection. Furthermore, the identification of one of these sequences (PY05828 in Sz5, Table 6.2) as being unspliced and the fact that nothing was observed in the -RT lanes gave cause for concern.

It was thought that there could be three possible explanations for these findings:

1) The sequences were from unprocessed mRNA

Experimental evidence seems to suggest otherwise:

- No partially processed mRNA had ever been discovered through sequencing of a large number of RACE or RT-PCR products crossing the introns.

- A band shift consistent with the splicing out of the second intron was observed in a pool of reverse transcribed cDNA (Fig. 3.1 chapter III).

However, it cannot be excluded that a few single cells had actually been processed at the exact time of mRNA synthesis. It is also possible that the unspliced sequence is from a subpopulation of transcripts. This is partly supported of the recent finding of unspliced *var* transcripts (Duffy et al., 2005).

Two single schizonts were found to contain one unspliced sequence each although the controls were negative (C5 and C6, Fig. 6.2 b and Table 6.1), when an exon 2/exon 3 primer pair was used. In addition, several individual PCR reactions from these cells reproducibly gave rise to strong intensity bands of the same size (not shown). Two different PCR reactions were also performed on the seven trophozoites, and sequencing revealed the same sequences. This indicated that the sequences originated from amplified material, as only one genomic copy of these repeatedly detected sequences is present in the uni-nuclear trophozoite.

2) Contamination had occurred

The amplicons could originate from contamination. This is not thought to be very likely for the following reasons:

- Products were only observed in the +RT lanes and not in the –RT or control lanes, which would be expected if this was caused by contamination.

- As a number of cells contained the same sequences, but no cells were found to contain them all, the contamination would have had to occur in a number of cases and each time only affects a few cells. For instance, the PCR for the schizonts was performed months before the PCR for the trophozoites and in between new SuperSMART kit, primers and PCR reagents and pipettes had been used. None of these sequences had been observed in any other sequencing occurring in the lab either before or after. Despite this, two identical pairs of sequences were detected in both schizonts and trophozoites, where no physical link in materials or reagents could be established.

3) The SuperSMART kit had amplified genomic sequences.

The possibility that the SuperSMART kit had amplified genomic sequences was thought to be the most likely explanation:

- Amplification of genomic sequences by the SuperSMART kit could explain the absence of products in the control lanes. After having divided each single cell into an experimental (+RT) and control (-RT) sample, only the experimental sample was put through the first strand synthesis reaction. At that time it was thought that any residual oligo dT primer would be removed by the column purification step (see materials and methods). However, in retrospect, a relatively large amount of oligo dT primer (7 μl of 12 μM oligo dT primer) was used in the first strand synthesis. If just a minute amount of this primer survived the column purification step and is present during the LD-PCR it could potentially anneal to the A/T rich genomic sequences. As this primer contains a tag sequence matching to the primers used in the LD-PCR (see materials and methods), this could lead to amplification of genomic material.

- The detection (Fig. 6.2 b) of strong bands containing unspliced sequences compared to the weak bands in the (-RT) controls indicates that a higher amplicon number is present in the (+RT) experimental samples than in the (-RT) controls, and this was not thought to originate from cDNA. Furthermore these bands were reproducible in individual PCR reactions.

- Co-detection of both spliced and unspliced sequences in a single cell. C8 (See Table 6.1) was found to contain both spliced and unspliced sequences. This would indicate that at a very low level of RNA and in an A/T rich organism, competition between LD-PCR amplification of both cDNA and gDNA could occur when using this kit.

Had both the +RT and – RT samples been processed identically during the first strand synthesis, PCR products of equal intensity would probably have been seen in both when running the reaction of a gel. However, as it is sometimes practice to have the –RT lane as a water control during reverse transcription this would clearly not have been sufficient in this application, even though the –RT samples went through the LD-PCR. This technology has been used successfully on other applications during this study, and in all cases, only spliced transcripts have been detected. But in all these cases, the starting material has been RNA extracted on a large scale from blood-stage infections and not single cells. In two studies of transcription in the hepatocytic stages (Sacci et al., 2005 and Gruner et al., 2005), the SMART technology was also used to amplify this limited amount of starting material. In one of these studies, detection of both spliced and unspliced sequences as well as minus strand sequences were reported (Gruner et al., 2005). Whether the detection of proposed minus strand transcripts and unspliced sequences originates from cDNA or gDNA is an open question, however the study performed here suggest that caution should be observed when using this technology on *Plasmodium*. Because of these concerns, only the four cells containing spliced transcripts were considered.

From these cells, it was found that, with a theoretical probability of at least 60%, that a maximum of two *yir* transcripts are present in a single infected RBC at the Schizont stage.

Chapter VII

Characterisation of *yir* UTRs

7.1 Introduction

Transcription initiation of eukaryotic genes occurs at various distances upstream of the initiation codon while termination happens some distance downstream of the stop codon, resulting in untranslated regions (UTRs) surrounding the coding sequence. The length of the UTR varies from gene to gene. On average, the 5′UTR on *Plasmodium* transcripts is 346 nt (Watanabe et al., 2002). This is slightly longer than the average 5′ UTR lengths observed for seven different taxa and which ranged between 103 nt and 221 nt (Pesole et al., 2001). There is a clear inverse relationship between the GC content of an organism and the length of its 5′UTR (Pesole et al., 2001) and this is thought to reflect the higher gene density, and consequently smaller transcription units, in GC rich organisms (Pesole et al., 2001).

Mapping the transcription initiation site is a starting point for identifying possible upstream regulatory *cis*-motifs, which in several eukaryotes can be located several Kb away, whereas the regions in the immediate vicinity of the transcription initiation site gives information of the motifs constituting the core-promoter, where the RNA polymerase II complex assemble. In *Plasmodium*, several genes seem to contain multiple transcription initiation sites (Watanabe et al., 2002), and the mapping of these sites has allowed for identification of upstream *cis*-elements for the *hsp86*, *calmodulin*, *cdp-diacylglycerol synthase* and *msp-2* genes (Militello et al., 2004, Polson et al., 2005, Osta et al., 2002 and Wickham et al., 2003). For some of these, *trans*-factor binding to these elements were shown, although these *trans*-factors were not identified.

For the multigene families, there could be different *cis*-elements interacting with different *trans*-factors for groups of these, and this could allow the parasite to differentially regulate these groups of genes. For the *var* genes in *P.falciparum* three major sets of 5′ intergenic regions, called upsA, upsB and upsC and two intermediate 5′ intergenic regions (upsB/A and upsB/C) are located in front of *var* genes located in distinct chromosomal regions and with distinct domain encoding capabilities (Voss et al., 2000, Gardner et al., 2002 and Lavstsen et al., 2003). Indeed, the upsB and upsC intergenic regions (located in subtelomeric and central chromosomal regions respectively) interacted differentially with nuclear factors at specific time points during an infection, and this correlated with the cessation of transcription (Voss et al., 2000).

7.1.1 Objectives

To investigate where transcription initiates and terminates for *yir* genes, 5′ and 3′ RACE were performed (7.2.1 and 7.2.2). To identify conserved motifs and relate these to the transcription initation and termination sites, intergenic regions were analysed through MEME analysis (7.2.3). One motif was found to be located immediately upstream of the 5′UTR (7.2.4). If *yir* genes are differentially transcribed, the conserved motifs could be distributed differently among the supergroups. Therefore, the *yir* supergroup distribution of the conserved motifs was investigated (7.2.5). As different sets of intergenic regions could confer differential regulation, the phylogeny of the 5′and 3′ *yir* intergenic regions was investigated (7.2.6 and 7.2.7)

7.2 Results

7.2.1 Experimental mapping of the 5′ UTRs

To map the UTRs surrounding the *yir* genes, 5′ and 3′ RACE were performed using the SMART RACE™ kit. RACE was performed using different primers (shown in

Fig. 3.1 a, Chapter III) and the resulting products were analysed using agarose gel electrophoresis (shown in Fig. 3.1 b, Chapter III). As was seen, only two of the primers (forward and reverse located in the same region in the third exon) yielded any products. In the 5′ RACE reaction, a single band of around 1.7 Kbp was detected while a more smeared product of >800 bp was observed in the 3′ RACE reaction. Several clones from each reaction were investigated by sequencing in both forward and reverse directions. For the longest 5′ RACE product additional primers based on the obtained sequence were used to ensure complete coverage of the whole insert. The gene identity of the 5′ RACE sequences was established by BLASTN analysis with sequences within the ORF against the annotated genes at www.tigr.org. Table 7.1 summarises the BLASTN analysis against the *yir* ORF to establish the gene identity. E value differences of at least two orders of magnitude between the best and second best BLASTN match were considered to be significant enough to assign the sequence to the *yir* gene suggested by the lowest scoring match.

The six analysed sequences (see Table 7.1) represented the longest sequences obtained from each of the six clones. In five of the six sequences, at least two orders of magnitude in E value difference existed between the best and next best BLASTN match. These were therefore assigned to the *yir* gene suggested by the best scoring match (lowest E value). The distances between the first nucleotide in the UTR and the translational start codon of the *yir* gene ranged from 426 to 772 nt (see Table 7.1). This suggested that transcription had initiated within a rather large 279 nt region relative to the ATG, and shows that *yir* gene transcription initiates several hundred nucleotides upstream of the ATG, generating long UTRs. None of the transcripts resembled each other in the region immediately before or after the first nucleotide. Five of the six identified genes belonged to supergroup 1 and one to supergroup 3, which reflects the findings in chapter III that more transcripts belonged to these two supergroups. However, it is worth mentioning here that the gene- specific primer used in the 5′ RACE was an exon 3 located primer that would not theoretically differentiate between genes in these two supergroups.

Table 7.1. BLASTN analysis of 5′ RACE products.

PY locus identifier[1]	Best/Next best match (E value)[2]	UTR length mapped on contig (nt)[3]	Group and Supergroup[4]
PY00842	8.6E-99/1.7E-92	619	G4/ SG1
PY04266	1.1E-101/3.1E-74	426	G6/ SG3
PY02298	4.8E-129/4.08E-108	543	G1/ SG1
PY00085	1.4E-137/3.2E-124	772	G4/ SG1
PY03195	5.3E-110/2.6E-100	493	G1/ SG1
PY03729(*)	3.8E-92/3.1E-92	601	G1/ SG1

1: Locus identifier for highest matching *yir* gene. 2: E values for best and next best matches. 3: Length (nt) of 5′UTR as mapped on contigs (first nucleotide to ATG). 4: Group/Supergroups identity of *yir* gene. In the case of (*), two almost identical *yir* genes were identified by BLASTN. However, the identity could be established through BLASTN with the 5′UTR, which only significantly matched one of these.

7.2.2 Experimental mapping of the 3′ UTRs

The 3′ RACE was performed with gene-specific primers located in the third exon (shown in Fig. 3.1 a, Chapter III). One additional clone was produced with a primer located in exon 2 (not shown), which gave information of the exon 2/exon 3 splice site. Due to the similarity between *yir* genes in this region, a BLASTN with the sequenced exon 3 would not allow for identification of a particular *yir* gene.

Therefore, identity was established from the 3′ intergenic regions themselves. A BLASTN-based approach was used, where the longest sequence from each of the clones was "BLASTed" against the genomic database at www.http//plasmodb.org and the difference in E value between the best and next best matches were investigated. Again, two orders of magnitude was considered sufficient to assign the identity of the 3′ UTR to the *yir* gene located on a particular contig. In all cases, the third exon was identified, and in one case the correctly spliced exon 2/exon 3 junction was also identified.

Table 7.2 BLASTN analysis of 3′ RACE products.

PY locus identifier[1]	Best/Next best match (E value)[2]	UTR length mapped on contig (nt)[3]	Group and Supergroup[4]	Poly A tail length (nt)[5]
PY03729	1.4E-90/1.2E-88	610	G1/SG1	ND
PY05586	4.4E-86/1.2E-81	630	G1/SG1	26
PY06920(*)	6.6E-87/2.5E-81	630	G2/SG1	ND
PY02541	5.3E-130/8.3E-91	543	G6/SG3	36
PY03398	2.9E-108/2.3E-78	615	G1/SG1	9

1: Locus identifier for highest matching *yir* gene. 2: E values for best and next best matches. 3: Length (nt) of 3′ UTR as mapped on contigs (last codon to transcript end). 4: Group/Supergroups identity of *yir* gene. 5: Length of the Poly A tail (in nt) that were in excess of the 31 As in the primer and where the primer tag was identified. The sequence (*) was obtained from another 3′RACE experiment (with an exon 2 primer) where a correctly spliced out intron 2 was observed.

At least two orders of magnitude difference between the best and next best matches between the sequence and contigs in the database (see Table 7.2). Therefore, the 3′

UTRs were assigned to the *yir* genes indicated in the table. It should be noted that PY03729 (contig MALPY01099) was also identified in the 5′ RACE, and therefore both UTRs for this gene were identified. The distances from the translational stop codons in exon 3 and the end of the 3′UTRs were measured on the contigs along with the groups and supergroup identity of associated genes.

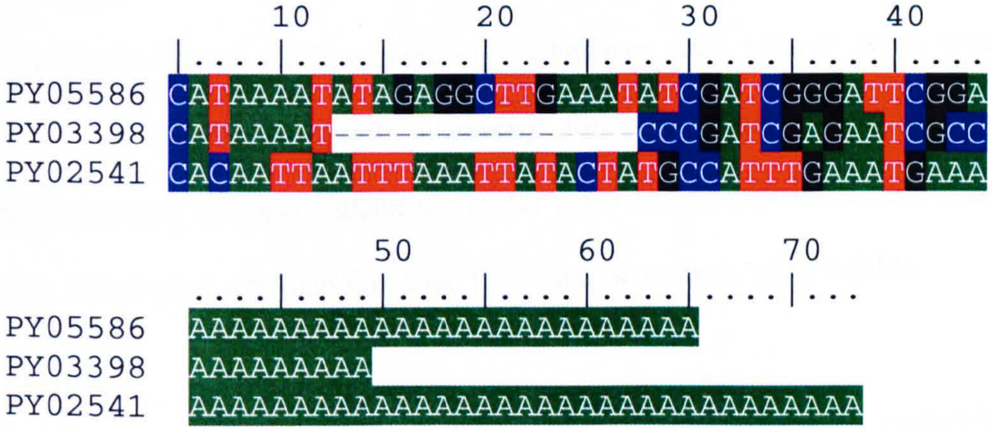
Genuine polyadenylated 3′ UTRs were identified in three instances. This was done by aligning the SMART 3′ RACE primer to the sequences. The sequence of this primer used in the reverse transcription was: 5′-AAGCAGTGGTATCAACGCAGAGTAC$(T)_{30}$V-N-3′ (N=A, C, G or T and V=A, G or C). After having identified the location of the primer in the sequences, any remaining stretch of As above the 31 which did not align to the contig, was considered part of the poly A tail.

Like the 5′ UTRs, the 3′ UTRs were several hundred nucleotides long. But, in contrast to the diversity in 5′ initiation sites, the termination of transcription seemed to have occurred within a narrower region of 87 nt with respect to the distance from the translational stop codon. As three of the 3′ UTRs were found to be polyadenylated, the sequences immediately before the polyadenylation, were investigated. From the alignment (Fig. 7.1) PY05586 and PY03398 had the same tri-CG repeat elements just before the polyadenylation site. PY03398 had a gap as can also be seen in the alignment, and this gap was also observed between the sequence and the contig PY03398 mapped to. PY02541′s 3′UTR did not resemble the two other UTRs in that particular region.

Figure 7.1

Detected 3′ ends of transcripts

Alignment between the ends of the three transcripts that ended in a stretch of A nucleotides that could not be accounted for by either the SMART primer or alignment to their respective contigs. In the alignment, the two first transcripts are polyadenylated after a similar region characterised by three CG repeats.



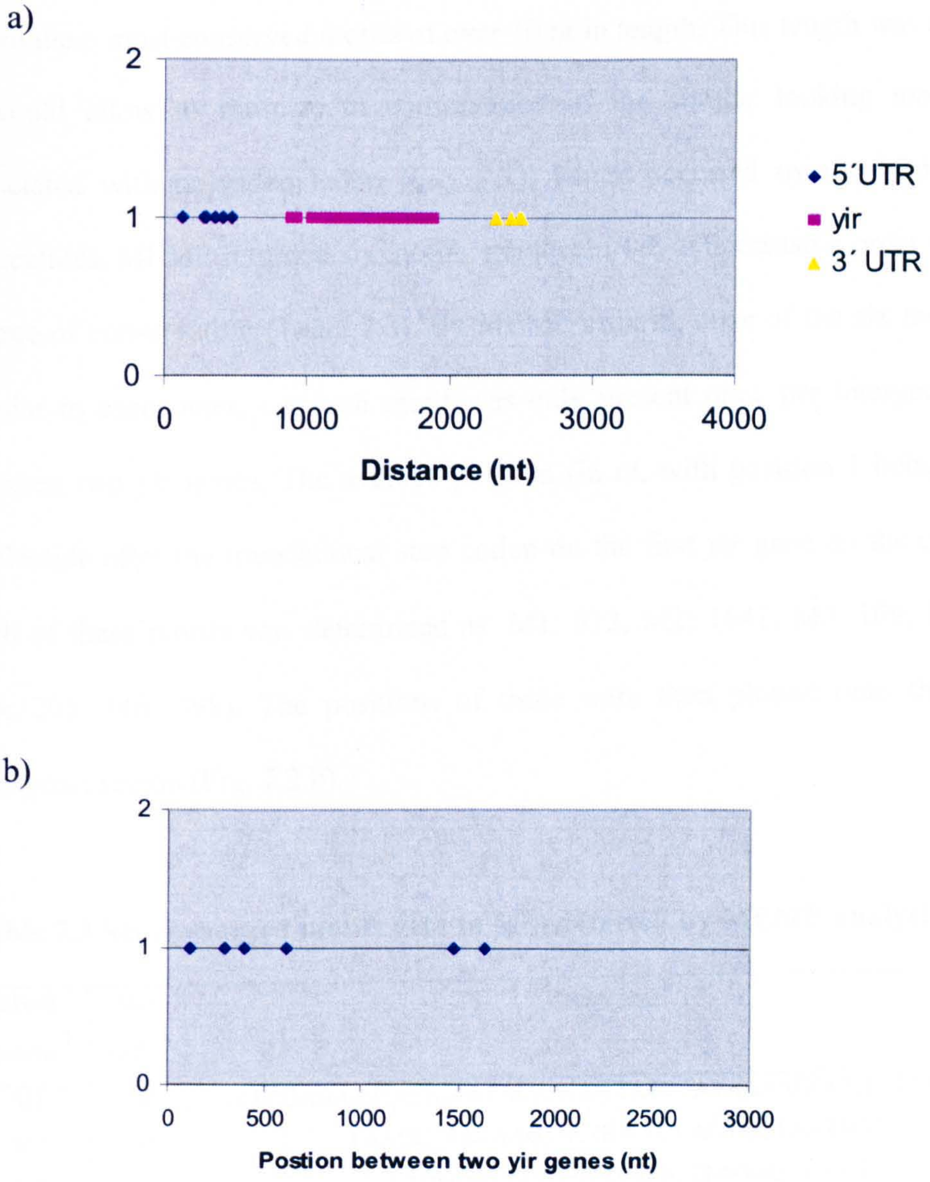
```
                 10        20        30        40
             |....|....|....|....|....|....|....|....
PY05586      CATAAAATATAGAGGCTTGAAATATCGATCGGGATTCGGA
PY03398      CATAAAAT---------------CCCGATCGAGAATCGCC
PY02541      CACAATTAATTTAAATTATACTATGCCATTTGAAATGAAA

                      50        60        70
              ....|....|....|....|....|....|....|...
PY05586       AAAAAAAAAAAAAAAAAAAAAAAAA
PY03398       AAAAAAAAA
PY02541       AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Figure 7.2

Lengths of intergenic regions and Location of conserved motifs

a) An arbitrary *yir* gene is shown, on which the distance from the ATG in exon 1 and the translational stop codon in exon 3 was kept as a reference point for the lengths of the 5′ and 3′ UTRs identified through sequencing. The longest 5′ UTR initiated 772 nt from the ATG and the two longest 3′ UTRs terminated 630 nt from the translational stop codon. The following sizes of 5′ UTRs were found: 772, 619, 601, 543, 493 and 426 (all in nt), and for the 3′ UTRs: 630, 630, 615, 543 and 449 nt.

b) MEME analysis was carried out on 24 intergenic regions separating 48 *yir* genes without any annotated ORF in between. The exact positions at which the *yir* ORF started and ended were manually identified, to ensure that only non-coding intergenic regions were analysed. The first position to the left of the figure corresponds to the first nucleotide after the translational stop codon in exon 3, and the last position to the right is immediately before the ATG in exon 1. The lengths of the intergenic regions varied from 1978-2835 nt, with an average length of 2403 nt (STDV 217 nt). This average intergenic region length was used to map the averaged locations of the six motifs (M1-M6) identified through the MEME analysis. The averaged locations of the six motifs were (positions indicate positions after the translational stop codon of the first *yir* gene on the contigs; left to right): M1:612, M2: 1641, M3: 109, M4: 1476, M5: 294. M6: 398 (nt). The sequences of the motifs are shown in Table 7.1.


a)
2
1
0
0
1000
2000
3000
4000
Distance (nt)
◆ 5´UTR
■ yir
▲ 3´ UTR
b)
2
1
0
0
500
1000
1500
2000
2500
3000
Postion between two yir genes (nt)

Using these criteria, intergenic regions of 48 *yir* genes were isolated from a total of 24 contigs. The lengths of the resulting intergenic regions varied from 1978-2835 nt, with an average length of 2403 nt (STDV 217 nt). MEME analysis was performed at http://meme.sdsc.edu/meme/website/meme.html, and the parameters were set to return the 6 most conserved motifs of over 10 nt in length. This length was chosen as it would allow to estimate to conservation of the similar looking motif found associated with polyadenylation (Fig. 7.1), which occurred over a region of 13 nucleotides. MEME returned six motifs, numbered 1-6 in decreasing order after their degree of conservation (Table 7.3). By MEME criteria, none of the six motifs were similar to each other, i.e. each motif was only present once per intergenic region between two *yir* genes. The average position (in nt, with position 1 being the first nucleotide after the translational stop codon on the first *yir* gene on the contigs) of each of these motifs was determined as: M1: 612, M2: 1641, M3: 109, M4: 1476, M5: 295, M6: 398). The positions of these were then plotted onto the average intergenic region (Fig. 7.2 b).

Table 7.3 Six conserved motifs (Ml to M6) detected by MEME analysis.

Motif number[1]	Size (nt)[2]	Sequence[3]
M1	50	GATCGAGAATCGACAATACAACGTTATCTATAAAAGGCGTTTTAGGCACC
M2	34	TAAGCTTATAAGTACGGGTCCAGTGCTAATTGTC
M3	29	GTGGGACCCATATTCGGGTTAGGGCTAAG
M4	50	AAGAGCAAATATATCTTATCTCTCTCCTCTCAAAGGGCAATATACCAACC
M5	41	TGAGTGTTCATGCCGATTTAATATGATTAAAATAAAATGTC
M6	50	CTTGATAACCCGGGGCTATATTGAATTATGCATATCACAATATGTTTCTT

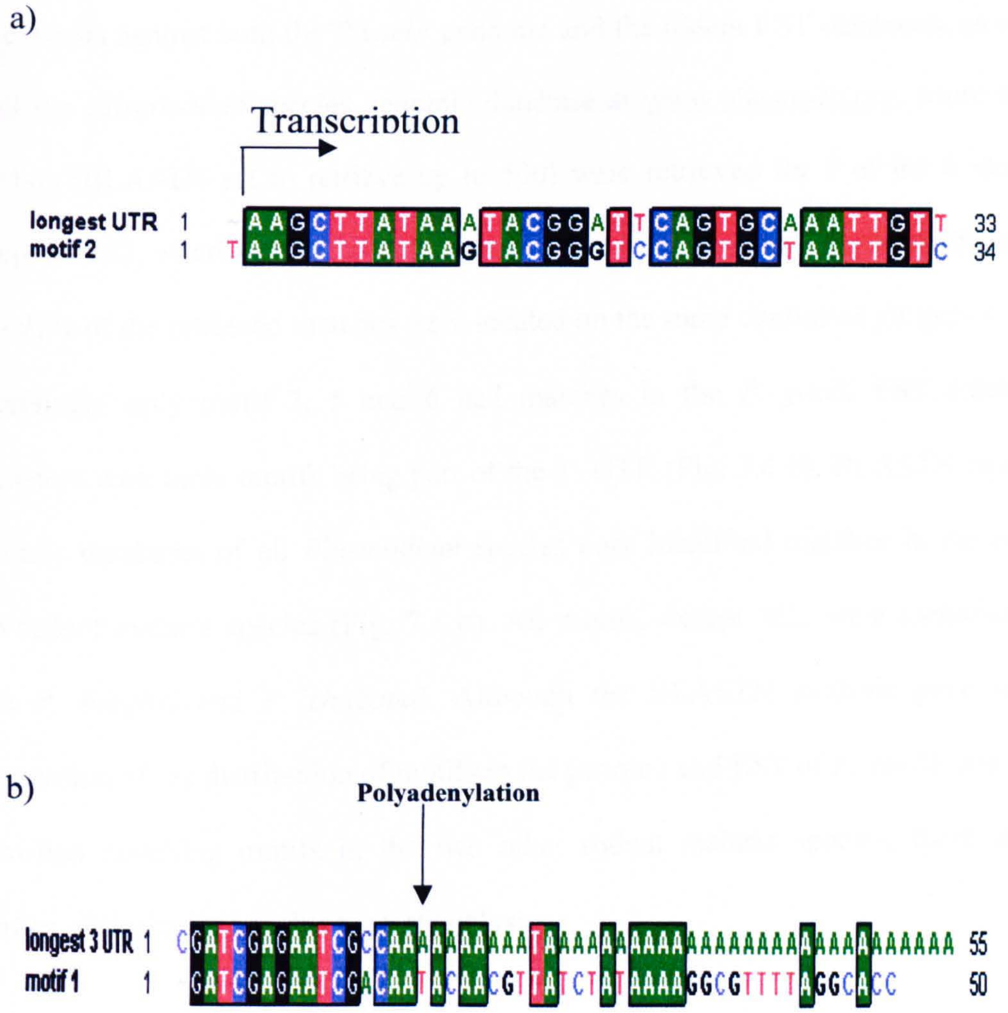
1: Motif number (Motif 2 is less conserved than Motif 1 etc.). 2: Size of motif in nt. 3: Motif sequence.

Figure 7.3

Matches between RACE results and identified conserved motifs

a) Shows the actual sequence from the longest 5′ UTR and its match with motif 2.

b) Shows the actual sequence from one of the longest 3′ UTRs and motif 1.


a)
Transcription
longest UTR 1 AAGCTTATAAATACGGATTCAGTGCAAATTGTT 33
motif 2 1 TAAGCTTATAAGTACGGGTCCAGTGCTAATTGTC 34
b)
Polyadenylation
longest 3 UTR 1 CGATCGAGAATCGCCAAAAAAAAAATAAAAAAAAAAAAAAAAAAAAAAAAAAAAA 55
motif 1 1 GATCGAGAATCGACAATACAACGTTATCTATAAAAGGCGTTTTAGGCACC 50

The start position of the longest 5′ UTR (Table 7.1) matched completely with motif 2 (Fig. 7.3 a), and two of the 3′ UTR transcripts ending with a genuine poly A (Fig. 7.1) tail matched completely with motif 1 (Fig. 7.3 b). To ensure that these motifs were unique to the *yir* intergenic regions, BLASTN was performed with each of these motifs against both the *P.yoelii* genomic and the rodent EST databases, as well as all the *Plasmodium* species genomic database at www.plasmodb.org. More than 100 hits (BLASTN set to retrieve up to 500) were retrieved for 5 of the 6 motifs except for M2, where only 12 matches could be found (Fig. 7.4 a). Importantly, more than 91% of the retrieved matches were located on the same contigs as *yir* genes. Not surprisingly only motif 3, 5 and 6 had matches in the *P. yoelii* EST database consistent with these motifs being part of the 3' UTR (Fig. 7.4 b). BLASTN against genomic databases of all *Plasmodium* species only identified matches in the other two rodent malaria species (Fig. 7.4 c). All motifs, except M2, were identified in both *P. berghei* and *P. chabaudi.* Although the BLASTN analysis gave some information of the distribution of motifs in the genome and EST of *P. yoelii*, and also identified matching motifs in the two other rodent malaria species, there are a number of limitations to this type of analysis:

1. Some contigs might be too short to contain both a *yir* gene and motifs. Although more than 91% of the contigs, where the motif was present, contained a *yir* gene, it is very likely that additional matches would have been identified in a chromosomal context.
2. Analysis of the EST database with various *yir* regions identified only a few matches, suggesting that the EST database is incomplete and contains a limited number of *yir* transcripts, which could indicate that *yir* is a low abundance transcript.

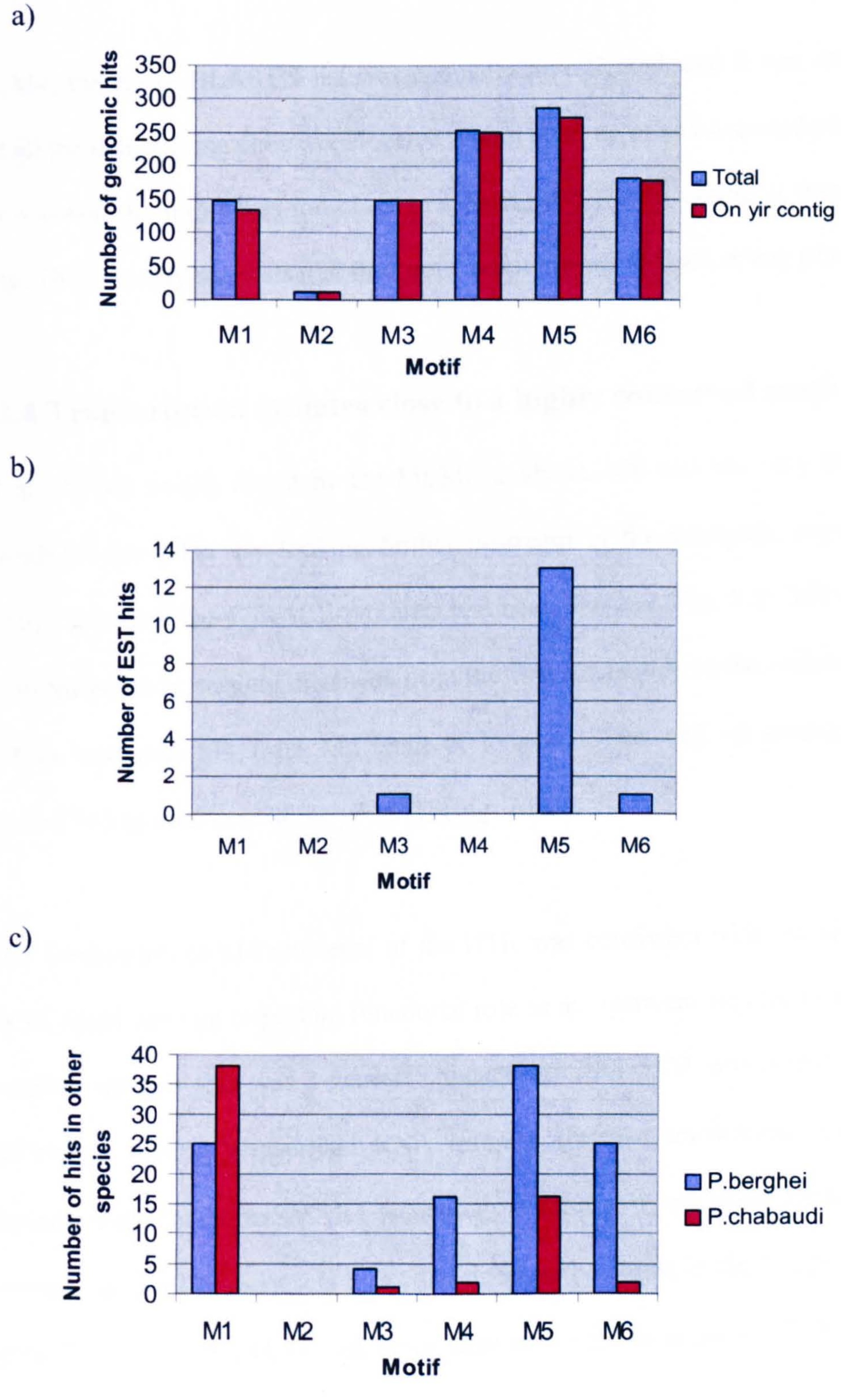
Figure 7.4

BLASTN analysis of motifs

a) Number of hits retrieved with a BLASTN with each of the six motifs (M1 to M6) against the *P. yoelii* genomic database at www.plasmodb.org. Location in the vicinity of a *yir* gene was determined manually and is also indicated.

b) Number of retrieved hits from a BLASTN analysis against the rodent EST database at www.plasmodb.org with each of the six motifs.

c) Number of retrieved hits against all *Plasmodium* genomic databases at www.plasmodb.org. Only hits to the two other rodent malaria species, *P.berghei* and *P.chabaudi* were retrieved from this analysis.


a)
Number of genomic hits
350
300
250
200
150
100
50
0
M1
M2
M3
M4
M5
M6
Motif
🟦 Total
🟥 On yir contig
b)
Number of EST hits
14
12
10
8
6
4
2
0
M1
M2
M3
M4
M5
M6
Motif
c)
Number of hits in other species
40
35
30
25
20
15
10
5
0
M1
M2
M3
M4
M5
M6
Motif
🟦 P.berghei
🟥 P.chabaudi

3. Since the motifs were rather small, it is possible that some sequence diversity exists, and that this diversity does not allow BLASTN retrieval.

For M4, the entire BLASTN list was manually investigated, and it was determined that all the detected matches were located within 1000 bp of an annotated *yir* gene. In a few cases, the motif was too close to a contig end for the contig to contain a *yir* gene. This strongly suggests that this motif is not present in front of any other gene.

7.2.4 Transcription initiates close to a highly conserved motif.

Of all the six motifs found by the MEME analysis, M4 was the only one located outside of the UTR. By looking further upstream in the intergenic region on the contigs to which the 5′ RACE products had been mapped (Fig. 7.5), M4 was found in all instances at varying distances from the first nucleotide in the transcripts. Only 138 nt separated M4 from M2 (start of longest UTR), and on average M4 was located 943 nt upstream of the *yir* ATG (Fig. 7.5).

The localisation of M4 upstream of the UTR was consistent with the idea that this motif could have an important functional role as an upstream regulatory element. To confirm whether this was a general characteristic, RT-PCR was performed on two differently reverse transcribed RNA samples (reverse transcribed using random hexamers and the SuperSMART technology respectively, see materials and methods) using a range of primers sets located at different positions in the 5' upstream region (Fig. 7.6 a, sets H and J), and these were also used in a control PCR reaction on gDNA.

Figure 7.5

Motif 4 and detected initiation sites

Motif 4 resembling sequences were identified and isolated from each of the contigs, to which the five sequenced 5′ UTRs were mapped. The distances between the initiation sites and the motif are shown. Also the distances between the initiation sites and the ATG in the first exon of the corresponding *yir* gene are shown. The average distance on the contigs between the ATG and motif 4 is indicated and was found to be 943 nt.

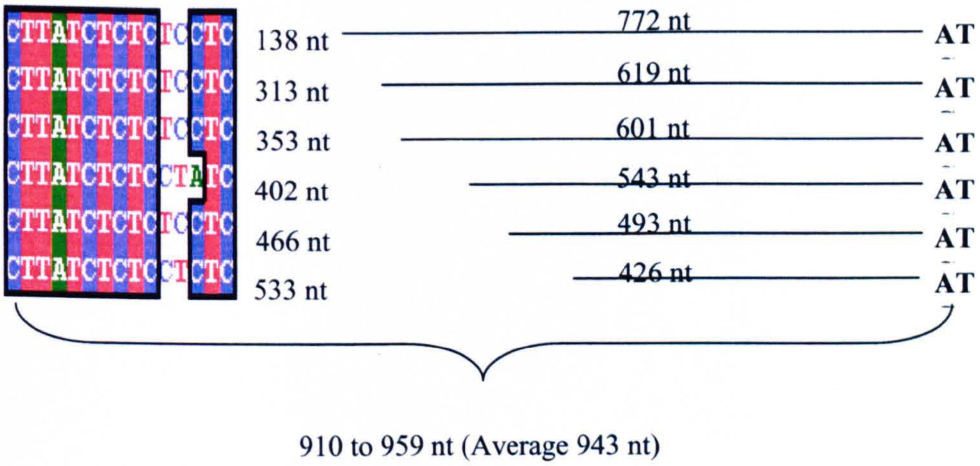

CTTATCTCTCTCCTC
CTTATCTCTCTCCTC
CTTATCTCTCTCCTC
CTTATCTCTCCTATC
CTTATCTCTCTCCTC
CTTATCTCTCCTCTC
138 nt
313 nt
353 nt
402 nt
466 nt
533 nt
772 nt
619 nt
601 nt
543 nt
493 nt
426 nt
AT
AT
AT
AT
AT
AT
910 to 959 nt (Average 943 nt)

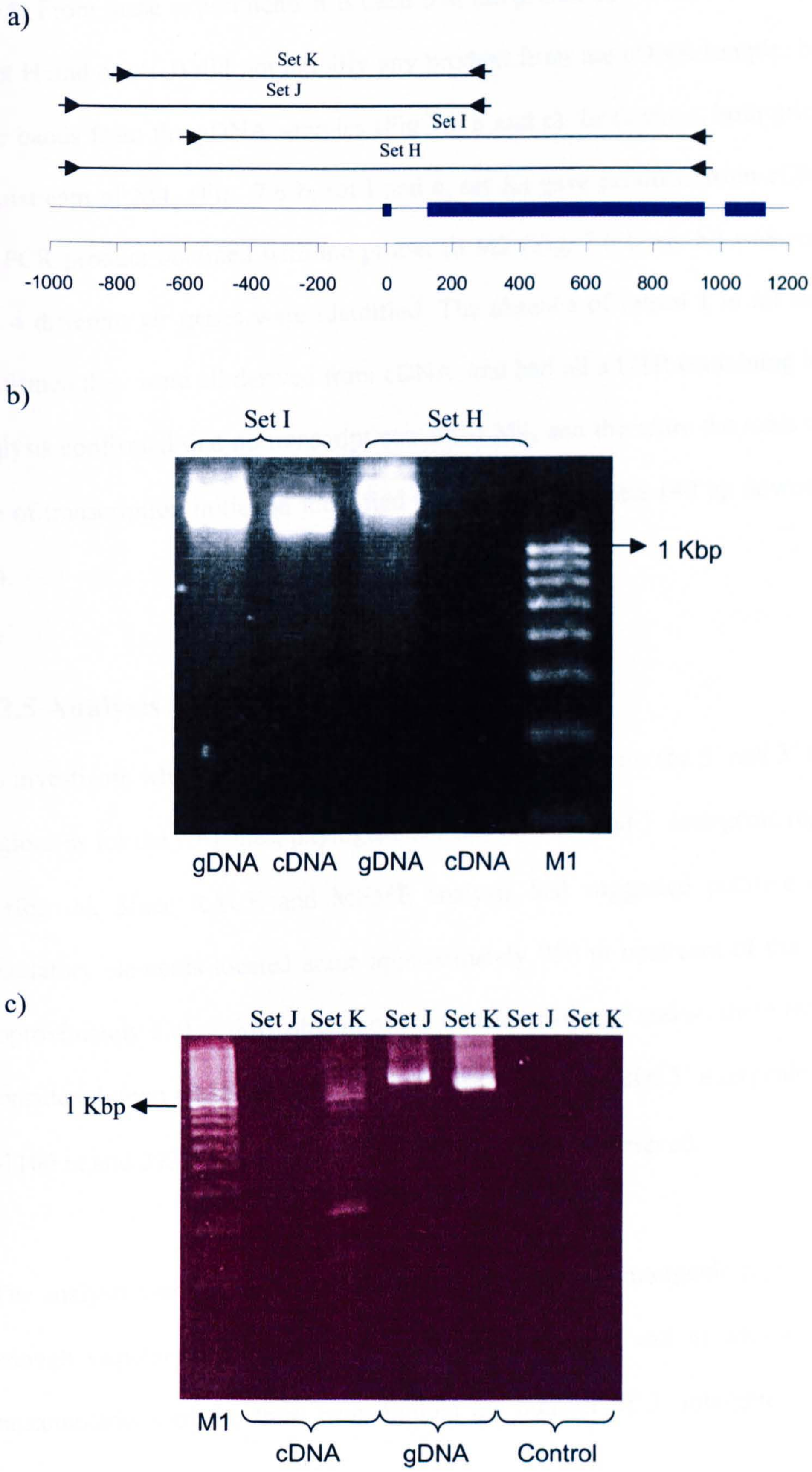
Figure 7.6

Mapping of the 5′ intergenic regions

a) Location of the used primer pairs. The expected size ranges of products amplified with the different primer sets were: set H: 1850-1950 bp (gDNA) and 1725-1825 bp (cDNA), set I: 1550-1650 bp (gDNA) and 1425-1525 bp (cDNA), Set J: 1200-1300 bp (gDNA) and 1075-1175 bp (cDNA), set K: 1050-1150 (gDNA) and 925-1025 bp (cDNA). The forward primer in set H and set J was located in M4, whereas the forward primer in set K was located in M2, and the forward set I primer some 300 bp downstream of M4.

b) PCR with primer set H and I on cDNA (generated with Superscript II reverse transcriptase, see materials and methods) and genomic DNA respectively. High intensity bands, much above 1 Kb, can be seen from both cDNA and genomic DNA with primer set I, whereas a strong intensity band can only be seen for genomic DNA with primer set H.

c) PCR with primer set J and K on cDNA (generated with the SuperSMART kit, see materials and methods) and genomic DNA. Bands of around 1 Kb were seen for both cDNA and genomic DNA with primer set K, whereas a band above 1 Kb was only seen from genomic DNA with primer set J. The set K cDNA product was cloned and sequenced.


a)
Set K
Set J
Set I
Set H
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
Set I
Set H
1 Kbp
gDNA
cDNA
gDNA
cDNA
M1
c)
Set J
Set K
Set J
Set K
Set J
Set K
1 Kbp
M1
cDNA
gDNA
Control

The strategy was to see if any primer set located in M4 amplified a product from cDNA. From these experiments it is clear that the primer set located in M4 (Fig. 7.6 b, set H and c, set J) did not amplify any product from the cDNA sample, but gave clear bands from the gDNA samples (Fig 7.6 b and c). In contrast, both primer sets downstream of M4 (Fig. 7.6 b, set I and c, set K) gave products from cDNA. The RT-PCR product obtained with the primer in M2 (Fig. 7.6 b, set K) was sequenced and 4 different *yir* genes were identified. The absence of intron 1 in all sequences confirmed they were all derived from cDNA, and had all a UTR containing M2. This analysis confirmed that no transcript contained M4, and therefore the most upstream site of transcription initiation identified was within M2 some 140 bp downstream of M4.

7.2.5 Analysis of the intergenic regions

To investigate whether a similar grouping could be found for the 5′ and 3′ intergenic regions as for the *yir* genes, phylogenetic analysis of 5` and 3′ intergenic regions was performed. Since RACE and MEME analysis had suggested putative conserved regulatory elements located some approximately 950 nt upstream of the ATG, and approximately 630 nt downstream of the translational stop codon, these lengths were considered upon retrieving intergenic regions. Therefore, 306 5′ intergenic regions of ~1100 nt and 292 3′intergenic regions of ~700 nt were retrieved.

The analysis was limited to this number as a number of intergenic regions lacked of enough sequence information on the relevant contigs, and in addition especially misannotations of the third exon limited the number of 3′ intergenic regions that could be retrieved.

From the alignment of the 5′ intergenic regions, M4 and M2 were easily identified visually, and after careful visual inspection it was found that they were each only located at one position (column) in the alignment. To visualize the conservation of M4 and M2, web logos were generated at http://weblogo.berkeley.edu/logo.cgi of all 306 sequences at the positions in the alignment, where these motifs were identified. M4 was found in all of the 306 5′ intergenic regions located at a very specific location in the alignment. Among all the 306 sequences, a conserved core motif: *CTTATCTCTC* was present. Some variation in the flanking sequences was observed, however this variation could not be distinctly correlated with the yir supergroups.

The M2 weblogo (Fig. 7.7 b) has two apparent characteristics: a TATAA like conserved motif followed by a GGG motif as indicated on the figure. The M2 weblogo appeared to be less conserved among the sequences, and therefore the 306 sequences were sorted after their resemblances to each other, and it was found that one conserved and one very diverse subset existed in that column. The very diverse subset consisted of 125 sequences with no overall resemblance to M2. The remaining 181 sequences were all very similar to M2, and a separate weblogo was generated from those (Fig. 7.7 c). The supergroup distribution of the 181 M2 motifs was analysed, and gave the following distribution: SG1: 55%, SG2: 45%, SG3: 63%, SG4: 38%, SG5: 43%. It is interesting to note that SG1 and SG3, where most transcripts were detected both by RT-PCR (see Chapter V) and RACE (this chapter),

Figure 7.7

Conservation of motif 4, 2 and 1 and supergroup distribution

From 306 aligned 5′ and 292 3' intergenic regions, the motifs were manually identified and the columns were isolated and realigned in ClustalX. The alignment from this was uploaded to the weblogo tool at www.weblogo.berkely.edu for analysis. This allowed for a visualisation of the degree of conservation. As the height of the letters in the weblogos represents their frequency within a column. The supergroup distribution of motifs was examined by scoring the number of each motif among intergenic regions belonging to the five supergroups and dividing it by the total number of analysed intergenic regions in each supergroup.

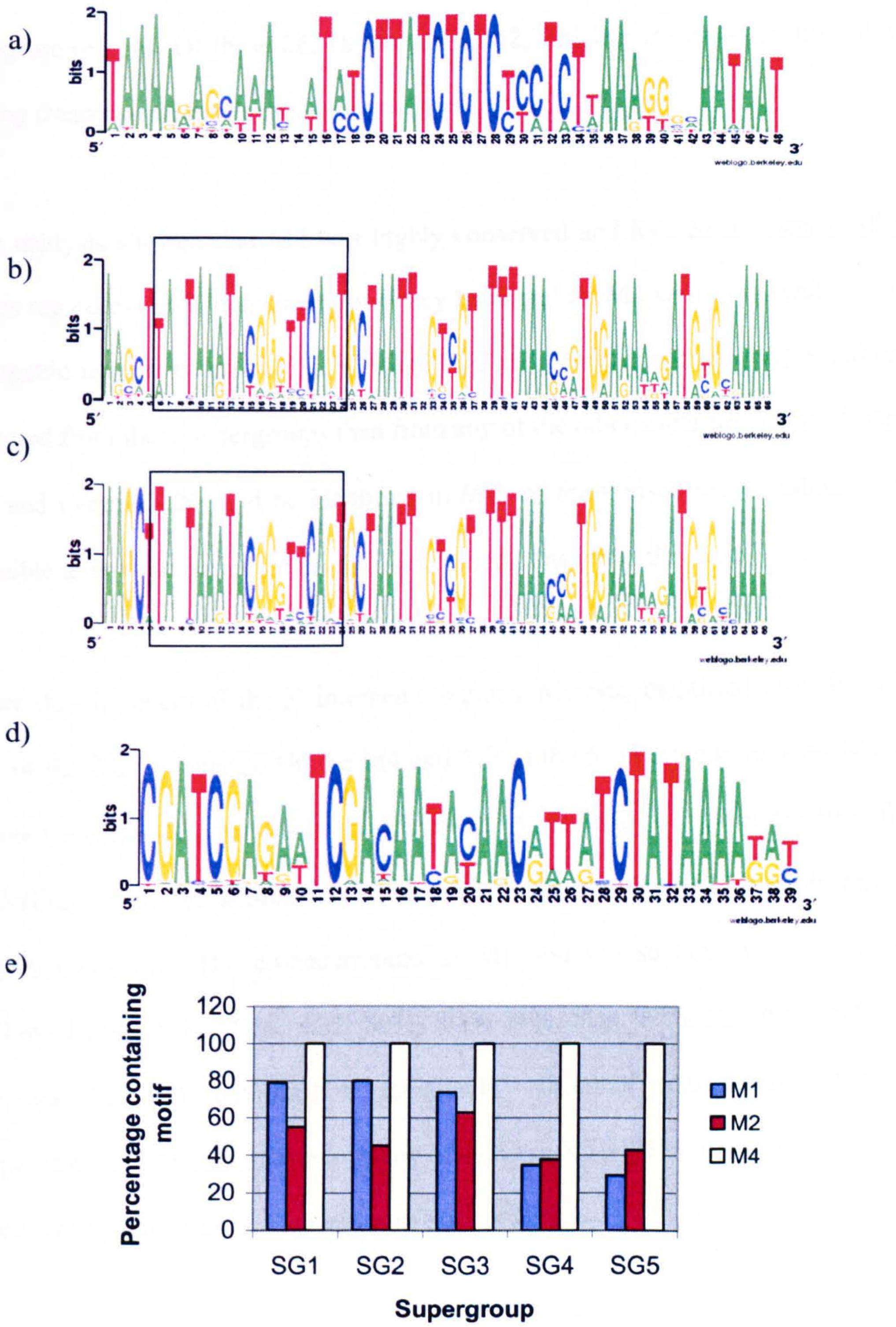
a) Shows the weblogo from all 306 5' intergenic regions in the column identified as containing motif 4.

b) Shows the weblogo from all 306 5′ intergenic sequences in the column identified as containing motif 2. In the boxed region is shown a TATAA like sequence and a C/G rich region.

c) Weblogo representation generated from the 181 sequences that were most similar to motif 2, as found by the sort function in BioEdit, which allows sorting by nucleotide frequency in selected columns from an alignment. In the boxed region is shown a TATAA like sequence and a C/G rich region.

d) Shows the weblogo from all 292 3′intergenic regions identified as containing motif 1.

e) Shows the percentages of intergenic regions belonging to SG1 to SG5 genes containing motif 1,2 and 4 respectively.


a)
b)
c)
d)
e)
bits
2
1
0
5′
3′
weblogo.berkeley.edu
Percentage containing motif
120
100
80
60
40
20
0
SG1
SG2
SG3
SG4
SG5
Supergroup
M1
M2
M4

had the highest proportion of M2, and the two most subtelomerically located supergroups; SG4 and SG5 (see Chapter IV) had the lowest proportion of M2. For 26 genes found to be transcribed in this study, it was possible to analyse their intergenic regions. Of these 26, 18 contained M2, which gives a distribution of M2 among transcribed genes a percentage of 69%.

This analysis showed that M4 was highly conserved and located in front of all *yir* genes regardless of which supergroup they belonged to. M2 was conserved in the 5′ intergenic regions of especially SG1 and SG3 *yir* genes. More transcripts had been detected from these supergroups than from any of the other supergroups (see Chapter V), and overall M2 could be identified in 69% of the transcripts, for which it was possible to analyse their associated intergenic regions in the database.

From the alignment of the 3′ intergenic regions, M1 was identified visually in 245 out of the 292 sequences. Unlike M4 and M2 in the 5′ intergenic regions, M1 was located at different positions in the alignment. A weblogo was generated from these 245 (Fig. 7.7 d) M1 similar motifs, and it can be seen that a *CGA* triple repeat is highly conserved. The co-occurrences of M1 and the supergroup *yir* gene they followed were: SG1: 79%, SG2: 80%, SG3: 74%, SG4: 35% and SG5: 29%. This showed that M1 exhibited a remarkably different distribution among the supergroups, with 2 to 3 times as many of genes in SG1-SG3 containing this motif in the 3′ intergenic regions, than genes in SG4 and SG5.

The distribution of M1, M2 and M4 (Fig. 7.7 e) shows that M4 is present in front of all analysed *yir* genes, M2 present in roughly 40-50% of the analysed *yir* intergenic regions, and there are some differences in the supergroup distribution of this motif, but it is not a truly universal motif. M1 is present in a remarkably higher proportion of SG1-SG3 *yir* 3 UTRs than SG4-SG5, however this motif is also not truly universal but is present in 29% to 80% of *yir* genes, depending on supergroup.

7.2.6 Phylogeny of the 5′ intergenic regions

Phylogenetic analysis was performed on the 306 5′ intergenic regions with Mega 2 and PAUP programmes (see materials and methods). From the bootstrapped UPGMA tree (Fig. 7.8 a) seven sets of sequences could be identified by their division at seven branch points (BP1-7). The reproducibility of these sets were tested in a ME tree (Fig. 7.8 b). This analysis shows that the overall clustering was identical except for a split of set 3 sequences into two different locations on the tree. Next, various bootstrap consensus levels were investigated, and at a 56% bootstrap value (Fig. 7.8 c) most of the outer branching points dividing the sets was still intact, except for set 1 and 2, which had developed into polytomies (e.g. many single branches, as low bootstrap value nodes disappeared). At a 95% bootstrap consensus level (Fig 7.8 d) all outer branch points had collapsed into polytomies except for the branch point for set 5. On an MP tree (Fig. 7.8 e and f) the same clustering into seven sets was observed, except that set 1 was split into two positions. The seven sets of 5′ intergenic sequences remained clustered to a high degree in all the tree building method used. The split of set 3 on the ME tree was not identified on the MP tree, although this tree in turn suggested a slightly different phylogeny for set 1. As these differences were not consistent between the ME and MP trees, and only appeared to separate a small number of sequences in the case of set 3 (compare Fig. 7.8 b and b), and for set 1 remained within a major cluster (Fig. 7.8 e and f), these sets were retained as they had been initially defined on the UPGMA tree (Fig. 7.8 a).

Figure 7.8

Phylogeny of the 5′ regions

a) Bootstrapped UPGMA tree of 306 5′ intergenic regions consisting of 1100 bp sequence upstream of the ATG. Seven visually identifiable branch points are marked by circles named according to the branch points as they occurs in a clockwise direction (BP1-7). All sequences inside these branch points were colour coded as seen in the legend.

b) Bootstrapped ME tree with sequences given the colour codes used in a).

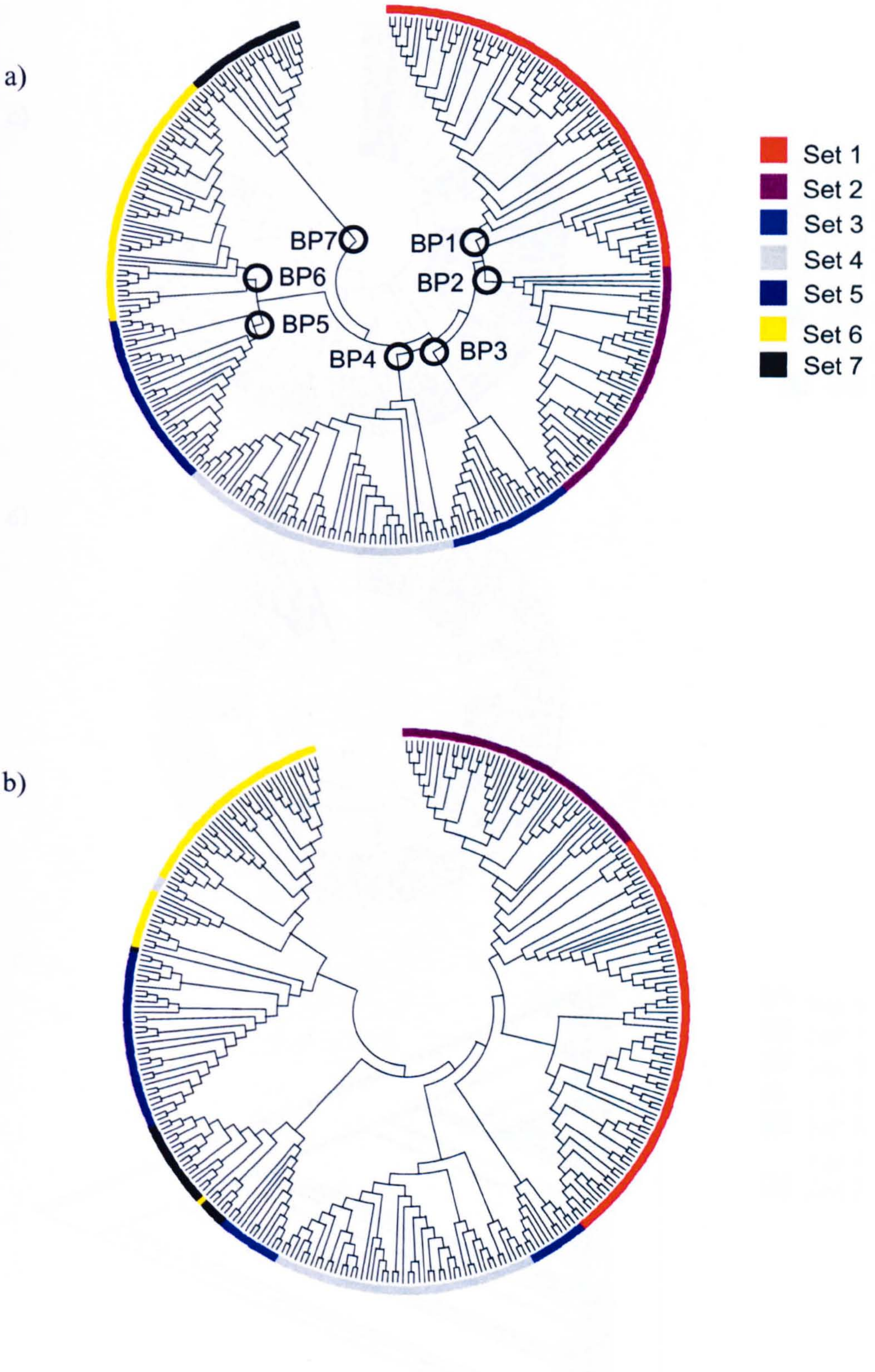
c) A 56% Bootstrap consensus UPGMA tree. Here, all branches with bootstrap values below 50% were collapsed.

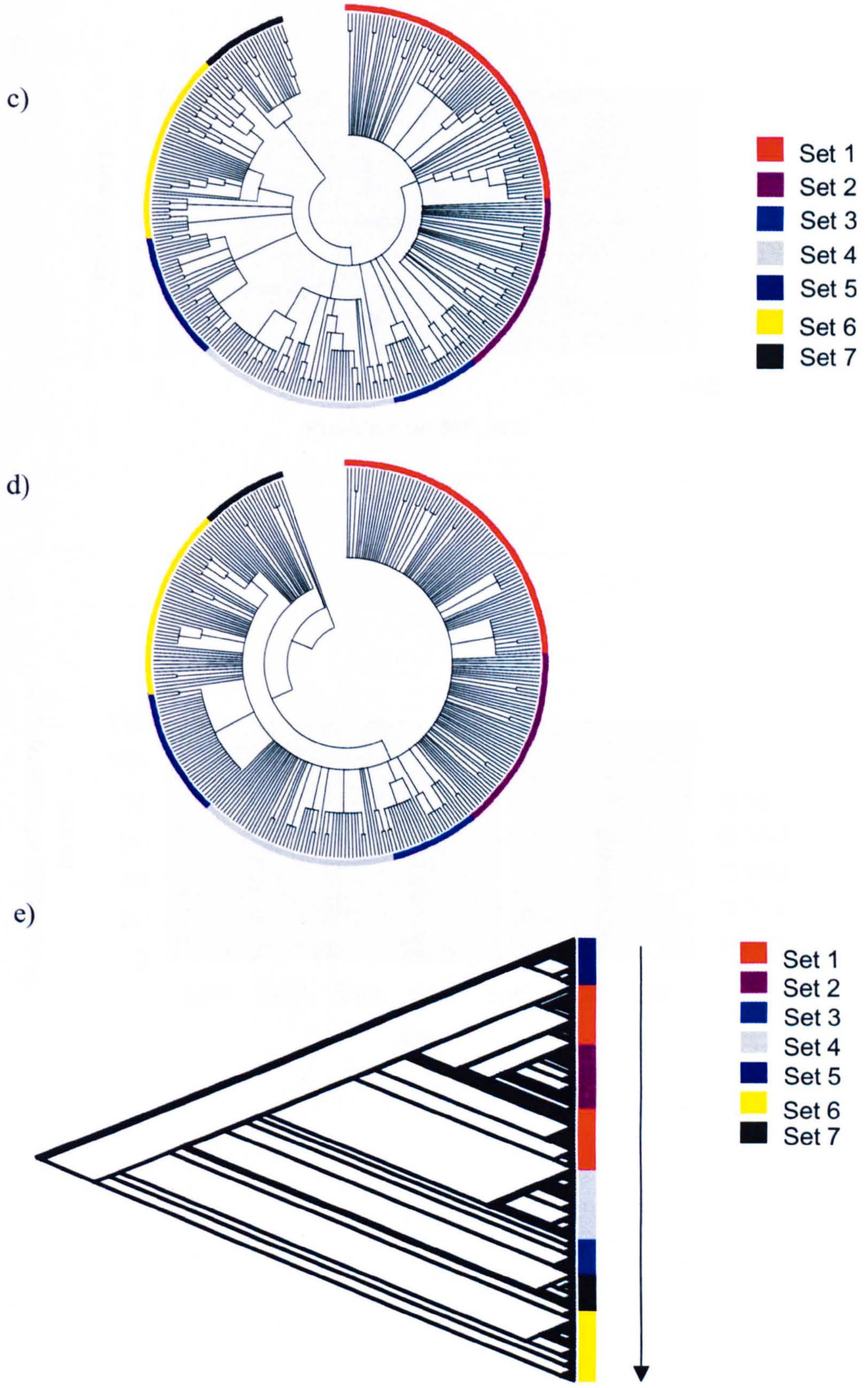
d) A 95% bootstrap consensus UPGMA tree. Here, all branches with bootstrap values below 95% were collapsed.

e) A Maximum Parsimony tree generated in PAUP, using the colour coded sequences identified. This colour coding only indicates the overall trend.

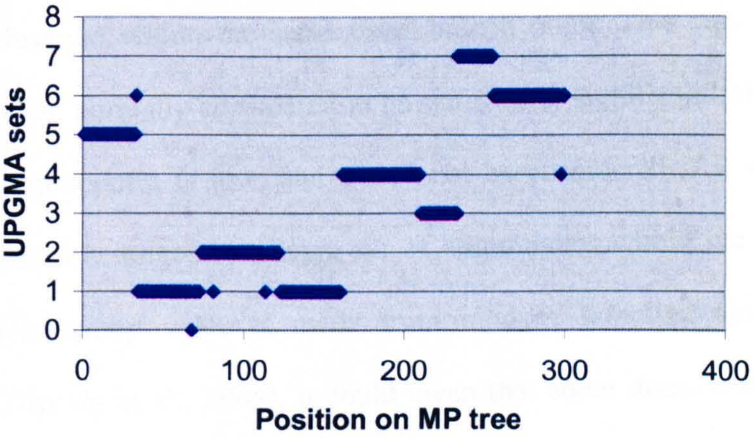
f) Sequences belonging to the seven UPGMA sets were plotted as they appeared in the direction of the arrow on the MP tree (figure e).

g) Percentages of sequences within each of the seven sets, which belonged to each of the five supergroups. From this, a clear co-occurrence of 5′ intergenic sets and *yir* gene supergroups was observed. As some sequences could not be assigned to a supergroup, the percentage does not add up to 100% in all cases.

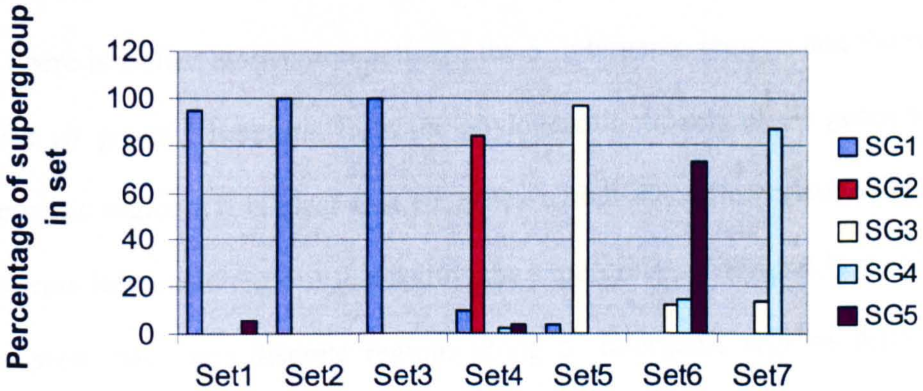

a)
BP1
BP2
BP3
BP4
BP5
BP6
BP7
Set 1
Set 2
Set 3
Set 4
Set 5
Set 6
Set 7
b)


c)
Set 1
Set 2
Set 3
Set 4
Set 5
Set 6
Set 7
d)
e)
Set 1
Set 2
Set 3
Set 4
Set 5
Set 6
Set 7

f)


UPGMA sets
8
7
6
5
4
3
2
1
0
0
100
200
300
400
Position on MP tree

g)


Percentage of supergroup in set
120
100
80
60
40
20
0
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set
SG1
SG2
SG3
SG4
SG5

The bootstrap consensus trees showed that, at 95% bootstrap values, almost all sets had collapsed into polytomies, whereas they were kept in all but two cases at 56% bootstrap values, and the two sets that did collapse at this level (set 1 and set 2) were both still clustered within the same outer branch point. This was below the 95% bootstrap value, normally considered to be statistically significant for phylogenies. A 56% bootstrap support is low, and would not be considered if it was not because these sequences belonged to a large set of homologues within the same organism. Because a bootstrap value is made from random sampling along the existing sequences (Brown et al., 2004), it could mean that some discrete regions along the 5′intergenic regions differed, whereas large parts remained the same.

When matching the 5′intergenic regions sets to the corresponding supergroup of *yir* genes they were located in front of, a very clear co-occurrence can be seen (Fig. 7.8 g). There is a clear association between the 5′ intergenic regions and the supergroups of the *yir* genes. Therefore, from the phylogenetic models of *yir* genes and their 5′ intergenic regions, it is clear that *yir* genes within the different supergroups contain upstream intergenic regions that follow the supergroups. From multiple alignments it was seen that some discrete regions along 5′ intergenic regions differ in a way, which correlates with the supergroup they belong to, and this could explain why the 5′intergenic sets were not separated by higher bootstrap values.

7.2.7 Phylogeny of the 3′ intergenic regions

A total of 292 3′ intergenic regions, containing sequence information ~700 nt downstream of the translational stop codon were phylogenetically analysed using Mega 2 and PAUP programmes. Within a bootstrapped NJ tree (Fig. 7.9 a), two major distinct branch points (BP1 and BP2) could be identified.

Figure 7.9

Phylogeny of the 3′ intergenic regions

a) Bootstrapped NJ tree of 292 3′ intergenic regions consisting of 700 bp sequence downstream of the translational stop codon. On the tree, two major visually identifiable branch points are marked by circles named according to the branch points as they occurs in a clockwise direction (BP1-2). All sequences inside these branch points were colour coded as seen in the legend.

b) Bootstrapped ME tree with sequences given the colour codes used in a).

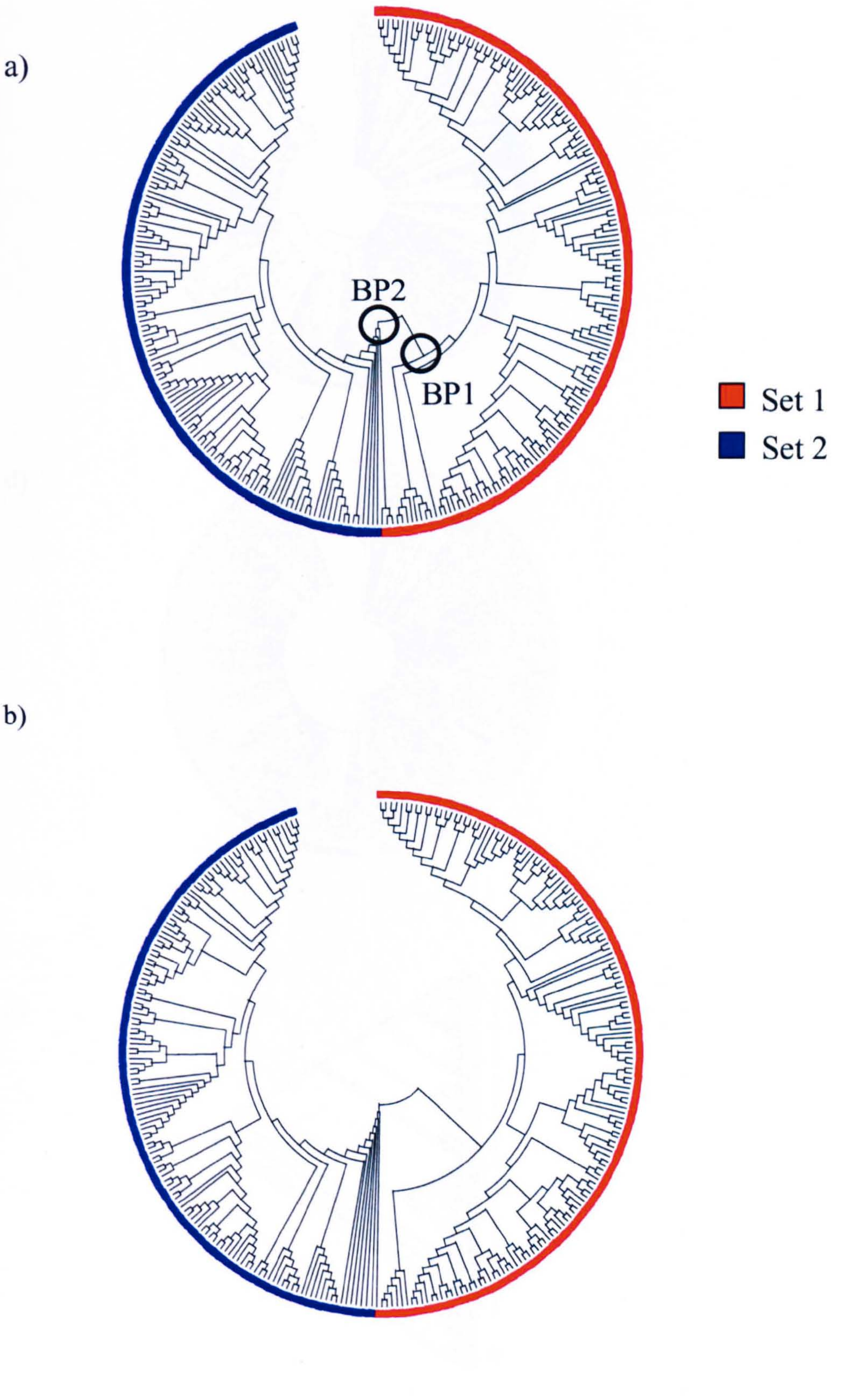
c) A 50% Bootstrap consensus NJ tree. Here, all branches with bootstrap values below 50% were collapsed.

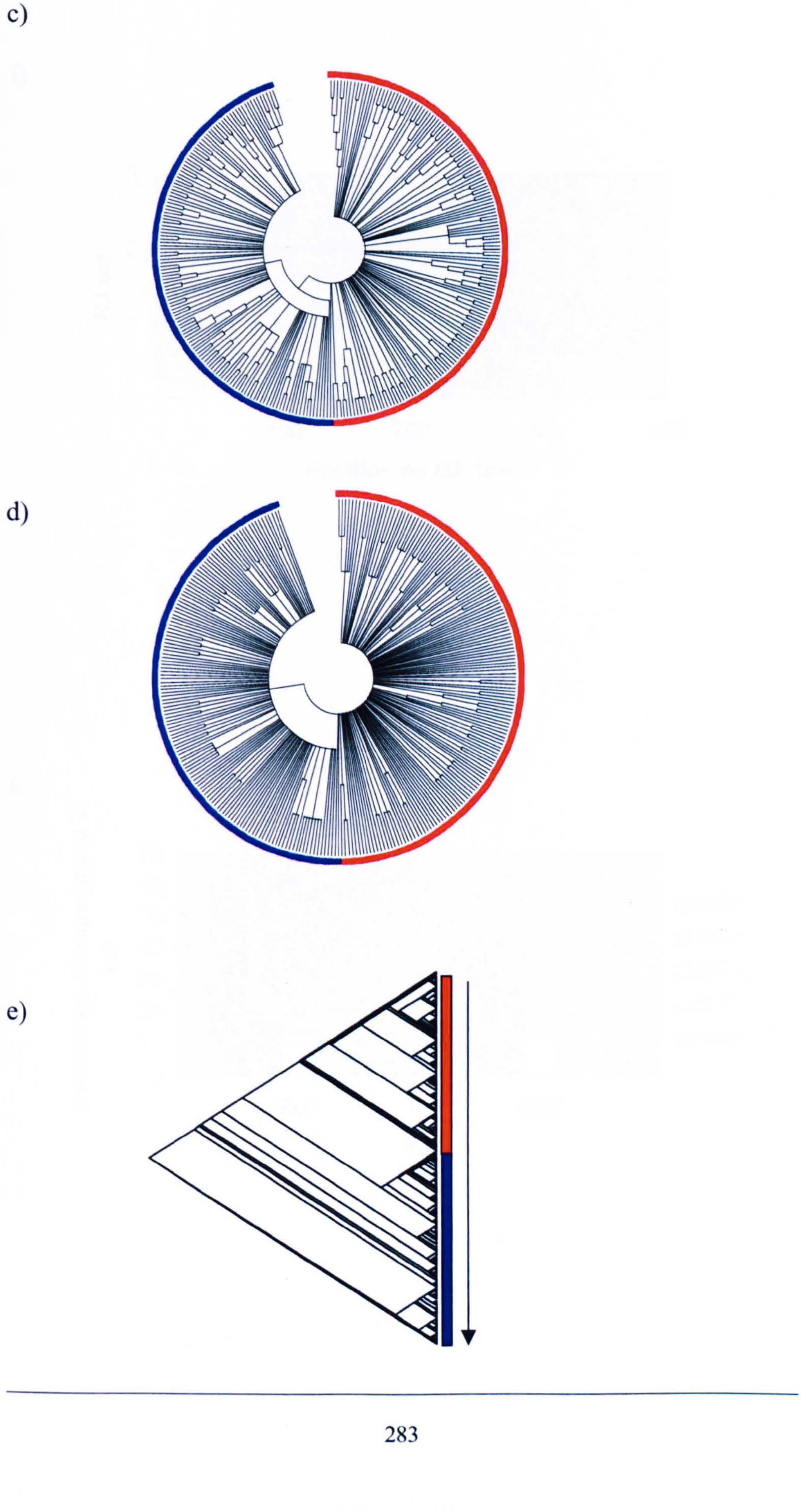
d) A 93% Bootstrap consensus NJ tree. Here, all branches with bootstrap values below 93% were collapsed.

e) A Maximum Parsimony tree generated in PAUP, using the colour coded sequences identified. This colour coding only indicates the overall trend.

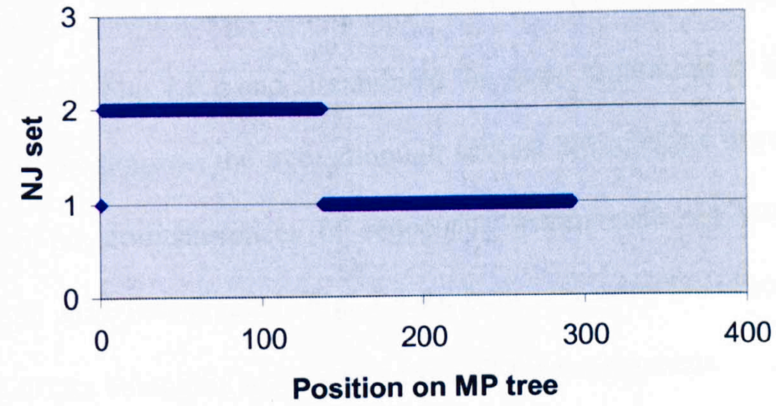
f) Sequences belonging to the seven NJ sets were plotted as they appeared in the direction of the arrow on the MP tree (figure e).

g) Percentages of sequences within each of the seven sets, which belonged to each of the five supergroups. No clear co-occurrence of 3′intergenic sets and supergroups could be observed as both sets contained an almost identical distribution of sequences from all supergroups. As some sequences could not be assigned to a supergroup, the percentage does not add up to 100% in all cases.

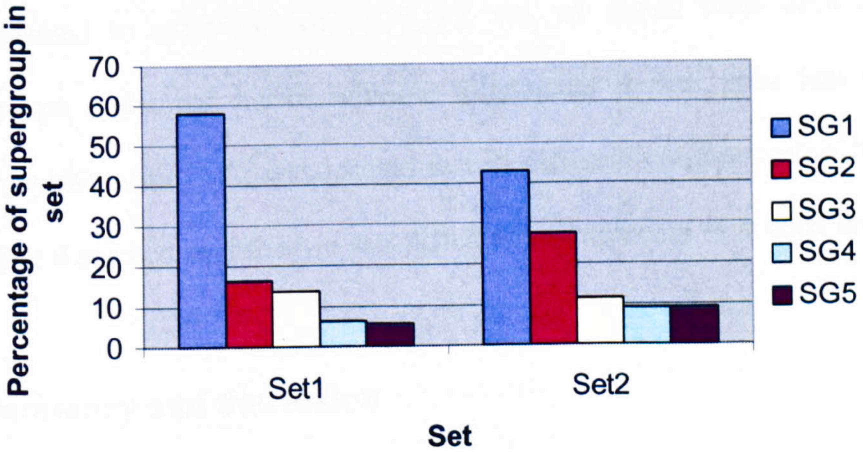

a)
BP2
BP1
🟥 Set 1
🟦 Set 2
b)


c)
d)
e)

f)


3
2
1
0
NJ set
0
100
200
300
400
Position on MP tree

g)


70
60
50
40
30
20
10
0
Percentage of supergroup in set
Set1
Set2
Set
SG1
SG2
SG3
SG4
SG5

In addition, several sets within each of these branch points were initially suggested from the phylogeny as well as some polytomies. A ME tree was generated (Fig. 7.9 b), and on this, the sequences within the two sets remained as seen on the NJ tree. When the bootstrap values of the NJ tree were investigated at 50% (7.9 c) and 93% (7.9 d), only set 1 and 2 were distinguishable from each other. A tree generated with the MP method (Fig. 7.9 e and f) exhibited the same separation of the sequences in two distinct locations on the tree, although several bifurcations were visible on this tree. When the co-occurrences of sequences within each set with the *yir* gene supergroups were investigated (Fig. 7.9 g), both sets were found to be present in a mixture 3 UTRs belonging to *yir* genes from all the supergroups.

The three-tree building methods all agreed in their clustering of the same two sets. Most of the possible sub-sets within the two major sets collapsed into polytomies at 50% bootstrap values, however the two sets were kept up till a 93% bootstrap consensus value. This bootstrap value is highly significant, however no particular order seemed to exist for which 3′UTR set, *yir* genes from each of the five supergroups contained. From multiple alignments it was seen that overall the polyadenylation site, M1, was located in two different columns some 100 nt apart. Therefore it seemed as if the two sets differed in the spacing of at least this motif.

7.3 Summary and discussion

In this chapter by 5′ and 3′ RACE established that *yir* transcription initiated at variable positions up to 772 nt upstream of the ATG and polyadenylation occurred some 630 nt downstream of the translational stop codon. Three conserved motifs were identified through MEME analysis, and two of these (M2 and M4) were located at, or close to, the transcription start site, and the third (M1) was found to coincide with polyadenylation. M4 was found to be located outside the transcribed region through RT-PCR analysis. The phylogenies of the intergenic regions were also examined, and the correlations between sets of intergenic sequences and the *yir* gene supergroups were investigated. This showed a clear correlation between 5′intergenic sets and *yir* supergroups, whereas no such correlation was found for the 3′ intergenic sets.

All the *yir* 5′ UTRs were larger (426 to 772 nt) than the average 346 nt for *P.falciparum* genes (Watanabe et al., 2002), and transcription seemed to initiate at various positions with no apparent similarity in their flanking regions. Despite the finding that diverse transcription initiation sites are not uncommon in *P.falciparum* (Watanabe et al., 2002), it cannot be excluded that some of these shorter 5′ UTRs originates from truncated transcripts, although the RNA used in this analysis was found to be of a high quality as estimated by gel electrophoresis (not shown). Sometimes, transcription start sites are measured through non-reverse transcriptase dependent methods. However, two RT-PCR studies validated that no transcript contained regions immediately upstream of this transcription initiation site, and it is not very likely that all transcripts in the two pools would have such strong secondary structures as to prevent reverse transcription of all *yir* transcripts present. For the 3′ UTRs, two transcripts ended in an identical region some 630 nt downstream of the translational stop codon, and were followed by a genuine poly A tail, whereas another polyadenylated transcript was detected 72 nt upstream of the two identical sites.

The three motifs (M1, M2 and M4), which coincided with transcriptional initiation and termination were all investigated for similarities to known regulatory elements in various databases (see materials and methods). M4 was found to be similar to a motif (consensus for *S.cervisiae* motif: *TCTCTCTCTCC*) interacting with a well-known pleiotropic regulatory factor, ABF1, in *Saccharomyces cerevisiae*. This regulatory factor has been found to bind to silencers, replication origins and centromeres (Springer et al., 1997, McBroom et al., 1994). M4 was also found to be located in front of *cir* and *bir* genes in *P.chabaudi* and *P.berghei* respectively. Since MEME designated this motif as M4, it meant that it was the fourth most conserved motif in this analysis. However, by BLASTN and multiple alignments, M4 was actually found to be much more conserved than both M1 and M2, and this probably reflects that the sample used for MEME analysis had been unrepresentative with regards to M1 and M2. M4 was not identified in front of any non-*yir* genes, and is therefore a defining characteristic of *yir* intergenic regions. The two RT-PCR analyses clearly showed that M4 was not a part of the transcribed region in two independent pools of mRNA.

If M4 is involved in transcriptional regulation, is it an upstream *cis* element or a part of the core promoter? The distance between M4 and the transcriptional initiation site for the longest 5′UTR is 138 nt. This distance is within the 81 to 186 nt distances earlier observed between *Plasmodium* core promoters and transcription initiation sites (Ruvalcaba-Salazar et al., 2005), however M4′s unique location in front of *yir* genes would suggest that this motif recruits *yir* specific *trans* factors and not just core promoter components.

M2 was not as conserved as M4, but was present in front of a higher than expected number of transcripts detected by RT-PCR (59% expected and 69% observed). Interestingly, more transcripts were detected from SG1 and SG3 (Chapter V) than any of the other supergroups, and these two supergroups also had a higher abundance of M2 in their intergenic regions. It is therefore possible that M2, maybe through its proximity to M4, is a more efficient transcription initiation site. However the absence of M2 in some intergenic regions before detected transcripts (31% of all detected transcripts) implies that other transcription initiation sites are used.

Two polyadenylation sites were identified through sequencing of the 3'RACE products. One, identical in two of the transcripts, coincided with a highly conserved triple-repeat *CG* motif found by MEME analysis. In both cases, M1 was flanked by the *CATAAA* consensus. This consensus was not included in M1 because it was separated from M1 by a short (10-20 nt) stretch of variable nucleotides. Interestingly, this consensus resembles the eukaryotic *AATAAA* box located some 10-30 nt upstream of a number of eukaryotic polyadenylation sites (Brown et al., 2002). M1 was highly conserved especially in SG1-SG3, however the experimental finding of another polyadenylation site, which bore no resemblance to M1, indicates that there are at least two polyadenylation sites in the 3 UTRs of *yir* genes.

It should be noted that neither M1, M2 nor M4 were found to be present in the intergenic regions of the long *yir* genes (see Chapter IV) and these genes are probably regulated differently than the remaining *yir* repertoire.

The 5' intergenic regions could be divided into seven sets (5'set 1-7), and the 3' intergenic regions could be divided into two major sets (3'set 1-2). The seven 5'sets were overall supported by bootstrap values up to 56%, while the two 3' sets were supported by a bootstrap value of 93%. The 5'intergenic sets are more similar than the 3' intergenic sets, which are reflected in the low bootstrap values. However, for the 5′ intergenic sets, a very clear co-occurrence of the supergroups and 5′ intergenic sets was found, while for the 3′ intergenic sets no such correlation was found. In addition, it appeared as if only discrete regions of the 5′ intergenic regions did differ in a way, which correlated with the supergrouping (Anticipating events somewhat, it should be noted here that a major difference between the sets was caused by the presence of an intron, but this will be discussed in detail in Chapter IX). The existence of distinct 5′sets in front of the *yir* genes resembles the upsA to upsC 5′intergenic sets located in front of the *var* genes (Voss et al., 2000 and Gardner et al., 2002). For the *var* genes, these sets relate to the chromosomal localisation of *var* genes, with upsA+upsB located in subtelomeric regions and upsC located in central chromosomal regions (Gardner et al., 2002). No *yir* genes were found on the same contigs as housekeeping genes (see Chapter IV), which suggested that the *yir* genes are not located in central chromosomal regions. However, the five *yir* supergroups exhibited a differential localisation on annotated subtelomeric and telomeric contigs (see Chapter IV). This more or less subtelomeric localisation is also reflected in the distribution of the yir 5′ sets and therefore resembles the ups-type and chromosomal localisation of the *var* genes. For the *var* genes, the ups type correlates with the domain encoding capacity of the associated *var* gene (Lavstsen et al., 2003), and certain theoretically possible domain encoding combinations are not found (Gardner et al., 2002 and Kraemer et al., 2003). This would suggest that *var* genes are kept within their respective groups through selection processes. One simple way of keeping this grouping is by physically separation of groups into discrete chromosomal regions, where inter-group recombinations are less frequent. Since genes on the same chromosome are not likely to recombine, it is possible that genes on different chromosomes located at the same distance from the telomere could recombine when the chromosomes align at the nuclear periphery (Freitas-Junior, 2002). In terms of gene regulation, the different *var* ups-types has been shown to interact differentially with nuclear factors (Voss et al., 2003), and recent studies have shown that central *var* genes with the upsC type was not associated with Sir2 whereas *var* genes with the upsA and upsB types were (Duraisingh et al., 2005). Whether the *yir* 5′ sets interact differentially with nuclear factors is an open and highly interesting question. The universality of M4 in front of all *yir* genes would suggest a universal capability among *yir* to recruit a *trans* factor involved in transcriptional initiation. However, the higher proportion of M2 in front of SG1 and SG3 genes and the overrepresentation of this motif in front of detected transcripts could suggest that SG1 and SG3 genes are transcribed more frequently due to the presence of M2. The 3′ intergenic sets transcended the *yir* supergrouping and the 5′intergenic sets. This could imply that the 3′ intergenic sets were not subjected to the same functional restrictions as the 5′intergenic sets. From manual inspection of these two sets, the polyadenylation motif, M1 was found to be located approximately 600 and 700 nt downstream of the translational stop codon in each set respectively. Since M1 was identified in 84% of analysed the 3′ intergenic regions, the most obvious difference between the two sets were that they had a different relative spacing with regards to the localisation of this motif. However, a lower proportion of SG4 and SG5 genes had M1 in their 3′ intergenic sets, and this implies that these supergroups used other polyadenylation sites. Since 3′ UTRs has been associated with translational regulation in *Plasmodium* (Golightly et al., 2000, Corredor et al., 2004 and Galinski et al., 2004) it is possible that this is also the case for the *yir* genes. It is therefore not possible to make any conclusions as to whether the 5′ intergenic regions and the supergrouping or the 3′ intergenic regions are functionally important. Both types could indicate different regulatory mechanisms.

Chapter VIII

Transfection experiments

8.1 Introduction

An unusual motif (M4, Chapter VII) was identified. M4 was located outside the transcribed region of *yir* mRNA molecules, 140 nt upstream of the transcription initiation site (M2, Chapter VII). In addition, M4 was found to be 100% conserved among all the yir 5′ intergenic regions available. No other conserved motifs were found upstream of M4 (MEME analysis, Chapter VII), suggesting that this motif could delimit a functional *yir* transcriptional unit. Furthermore, M4 was found to be similar to a motif interacting with a well-known regulatory factor, ABF1, in *Saccharomyces cerevisiae* (Springer et al., 1997, McBroom et al., 1994). Therefore, M4 could be involved in transcriptional regulation of *yir* genes, and this was investigated further.

8.1.1 Objectives

Details of cloning and transfection are described in materials and methods (See 2.9 and subsections, Chapter II), and these are therefore only briefly described here. To investigate if this motif was involved in *yir* gene regulation, a transfection vector was used for transient transfection experiments (8.2.1). A *yir* intergenic region (8.2.2) was cloned into the transfection vector, and a nested deletion of M4 was made (8.2.3). In an initial transfection, both expression and transcription was investigated (8.2.4), however a negative control was absent from this experiment. In the final transfection all relevant controls were included (8.2.5).

8.2 Results

8.2.1 Transfection vector

Two transfection constructs, *PTubGFP*M3 and *PDEFGFP*M3, were described earlier (See 2.9.1 and Fig. 2.4). These had been previously shown to work successfully in *P.berghei* transfections (Blandine Franke-Fayad et al 2004) (kind gift of Andy Waters, Leiden). The constructs contain *gfp* under control of either the *P.berghei* elongation factor (*Pbef*, Fig. 2.4 a), or the *P.berghei* tubulin α-II promoter (*Pbtub*α-II, Fig. 2.4 b). The *Pbef* promoter was found to drive a strong GFP expression in all stages (Blandine Franke-Fayad, 2004), while the *Pbtub*α-II promoter only drove GFP expression in male gametocytes (Blandine Franke-Fayad et al., 2004). In this study, the *Pbtub*α-II promoter was replaced by a *yir* intergenic region, while the construct containing the *Pbef* promoter was used as a positive transfection control. Transient transfections for episomal expression were performed under drug (pyrimethamine) cover.

GFP expression was measured by Fluorescence Activated Cell Sorting (FACS), which allows the user to collect measurements from a defined number of cells. Unlike measurements, where one average reading is performed per sample, FACS has the advantage of giving the distribution of readings within a population of individual cells, and in addition allows the user to gate by several criteria. For studies of *Plasmodium*, one of the gates measures the DNA content of each individual cell by pre-incubation with the DNA binding HOECHST due prior to the FACS analysis. As erythrocytes do not contain DNA, these are excluded from the analysis. It is also possible to distinguish different parasite stages based on their DNA content. In this study, emission from GFP was used to measure the activity of the transfection constructs. This vector produced a GFP protein with an excitation wavelength of 488 nm and emission of the green fluorescence was measured at 530 nm (Blandine Franke-Fayad, 2004).

8.2.2 Amplification of *yir* intergenic sequences

Yir intergenic regions were obtained by PCR with primers and analysed by sequencing as described (See 2.9.3). A 1019 bp *yir* intergenic region, called IR1019 was chosen for cloning.

IR1019 was highly similar (91-93% identity) to three contigs in the database (www.tigr.org), however there did appear to be some regions where IR1019 was more similar to one of the contigs than the other. Therefore an alignment with all three contigs was performed (see supplementary, S8.1), and from this it can be seen that within the first 100 nucleotides there are three gap regions, of which one (position 68 to 78 in the alignment) is consistently different between IR1019 and the three contigs. The two most likely explanations for this gap could be that either the best hit could not be identified in the database, due to contig ending, or that it was among the estimated 10% of the *P. yoelii* genomic DNA missing from the 5X coverage. It could also be that strain differences do exist. Since a 100% match could not be established, it was investigated how much IR1019 differed from the three best matching contigs, and it was found that 97% of the nucleotides in IR1019 matched to at least one of the three contigs. The region of IR1019 containing M4 (see position 110 to 126 in supplementary S8.1) was completely identical to the contigs. In addition, all sequences ends in the alignment immediately before three SG1 *yir* genes, of which one had earlier been found to be transcribed (Chapter VII).

Whether IR1019 represents a functional intergenic region was to be established during transfection experiments, but at this stage, its resemblance to an intergenic region from a gene that had been identified as a spliced transcript by 3′ RACE and the presence of a M4 motif at position 110 to 121 made it a likely candidate.

8.2.3 Cloning

In order to assess the function of M4, a construct containing M4 and one where it was deleted, was needed. In addition, a vector with no promoter was needed as a negative control. A schematic show how these three constructs were generated (Fig. 8.1 a to c and for a more detailed description see 2.9.4 to 2.9.6, Chapter II). By sequencing (See supplementary, S8.2) it was confirmed that all three constructs had been correctly inserted and that no nucleotide changes had occurred during the cloning procedures. These three constructs were called pIR1019 (+M4), pIR894 (-M4) and pTub (-) (negative control).

8.2.4 Transfection in *P.yoelii* 17X (Transfection 1).

Initially, it was tested whether pIR-1019 could drive GFP expression/*gfp* transcription. Electroporation and transfection procedures are described in materials and methods (See 2.9.6 to 2.9.9, Chapter II). Female BALB/c mice were injected with pIR1019 and pEF constructs along with a drug and mock transfection (electroporation performed only in PBS) controls. Parasite development in the infected animals was monitored and parasitaemia curves are shown (See supplementary, S8.3). FACS analysis was performed as described (See 2.10.1, Chapter II) after having bled the mice at the end of the second passage.

A high proportion (71 to 77%) of the iRBCs transfected with pEF, the positive control vector, was GFP positive (Fig. 8.2 a and b), showing that transfection had been successful and that the construct was functional in *Plasmodium yoelii.*

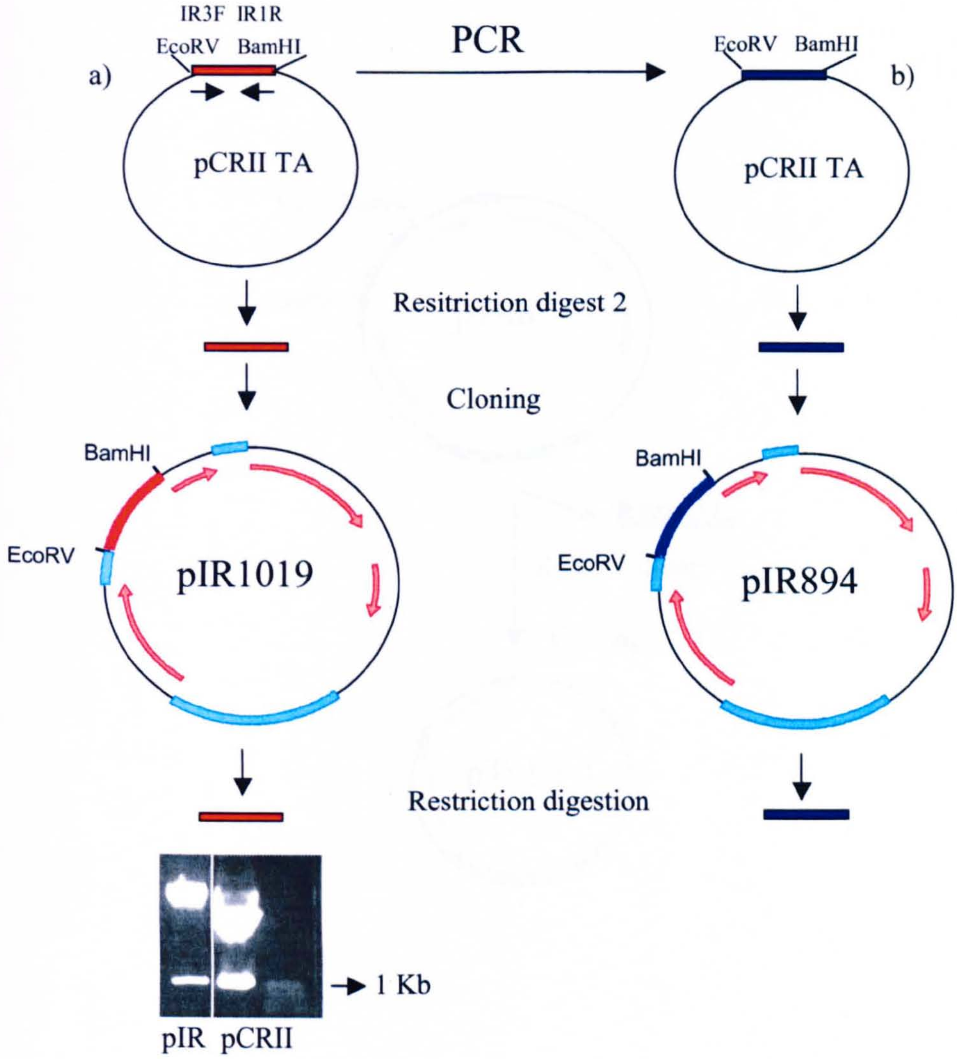
Figure 8.1

Construction of vectors

a) The pIR1019 plasmid was generated by first excising the 1019 bp IR fragment from the pCRII TA sequencing vector with EcoRV/BamHI digestion (restriction digest 1). This fragment was then isolated by gel electrophoresis and electroelution followed by phenol/chloroform extraction. A linearized transfection vector, from which the tubulin α-II promoter had been removed, was generated alongside. The IR1019 fragment was cloned into the EcoRV and BamHI sites of the linearized transfection vector by overnight ligation at 20 ° C using T4 ligase at various insert: vector ratios (1:5, 1:3, 1:1 and 3:1). Controls were included where T4 ligase or the IR1019 fragment was omitted. Successfully generated pIR1019 clones were identified by EcoRV/BamHI digestion (restriction digestion, example shown) alongside the pCRII TA vector containing the IR1019 fragment, both generating a fragment of 1019 bp (gel). (See also materials and methods).

b) Since no suitable restriction enzyme site existed in IR1019, a primer (IR 3F) was designed 125 bp after the start of IR1019 (excluding M4 from product) and used to generate the nested M4 deletion fragment, IR894. IR 3F contained an Eco RV tag and was used in combination with primer IR 1R that contained a BamHI tag. After cloning this 894 bp PCR fragment into the pCRII sequencing vector, it was cloned into a linearized transfection vector in accordance to the generation of pIR1918 (a). The resulting vector was called pIR894. (See also materials and methods).

c) The pTub (-) transfection vector control was generated by Eco RV and Bam HI digestion and removal of the excised tubulin α-II promoter fragment by gel electrophoresis. The overhang was filled with Klenow enzyme, and the vector was religated using T4 ligase as described above (a). Successful generation of pTub (-) clones were screened by insensitivity to EcoRV/BamHI digestion (not shown). (See also materials and methods).


IR3F IR1R
EcoRV BamHI
a)
pCRII TA
PCR
EcoRV BamHI
b)
pCRII TA
Resitriction digest 2
Cloning
BamHI
EcoRV
pIR1019
BamHI
EcoRV
pIR894
Restriction digestion
1 Kb
pIR pCRII

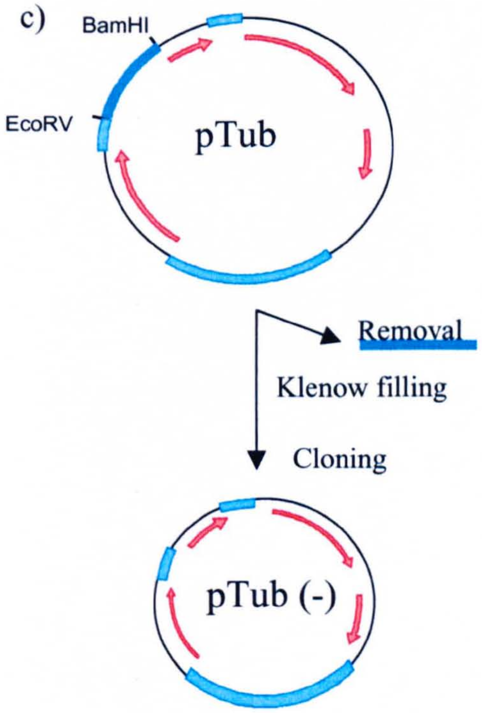

c)
BamHI
EcoRV
pTub
Removal
Klenow filling
Cloning
pTub (-)

Figure 8.2

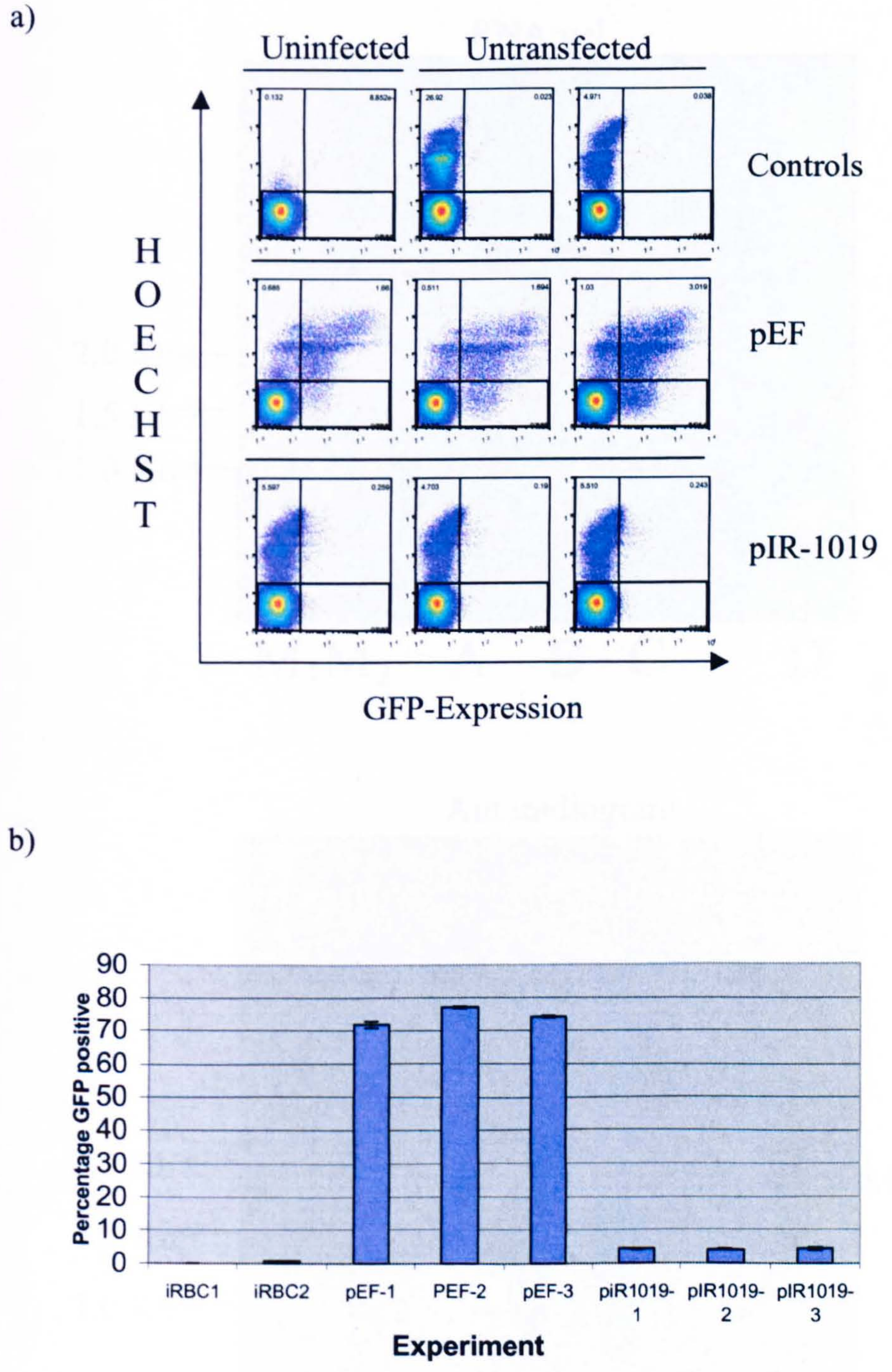
17X FACS and Northern blot

a) FACS analysis of 17X transfections. The FACS plots (from top to bottom) shows: Controls (only controls contain both uninfected erythrocytes and untransfected, but infected erythrocytes respectively), positive Control (pEF transfected and Pyrimethamine selected *P.yoelii* 17X parasites, see materials and methods), and Experimental samples (pIR1019 transfected and Pyrimethamine selected *P.yoelii* 17X parasites). The X axis (FL1) shows the GPF expression and the Y axis (FL4) shows the HOECHST staining intensity. A total of 50000 events were collected per sample, and all measurements were performed in duplicates.

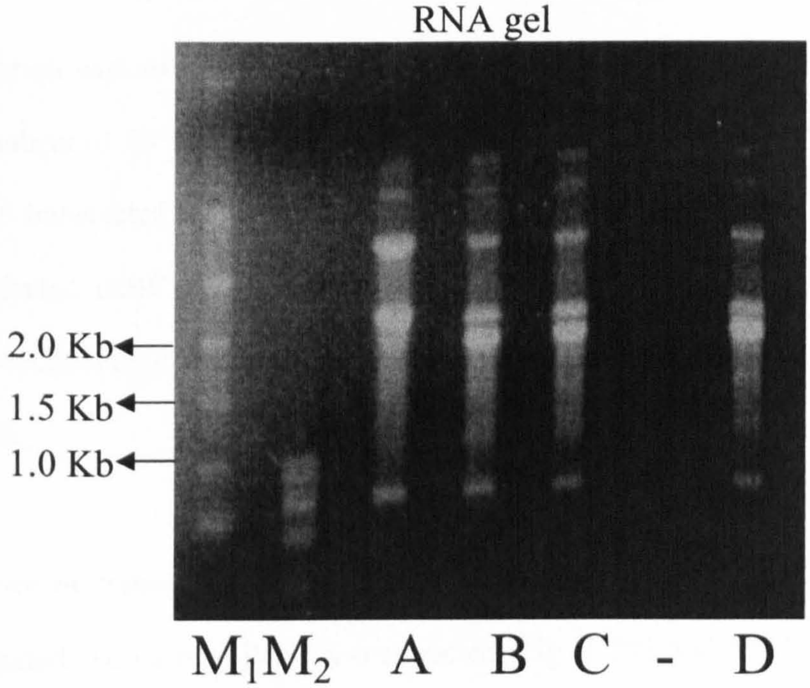
b) The percentages of double positive HOECHST/GFP events for the untransfected negative controls (iRBC 1-2), the positive controls (pEF1-3) and the experimental controls (pIR1019 1-3).

c) Equal amounts of RNA from iRBC 1, pEF 1&2 and pIR1019 3 was loaded on the gel in lanes A, B, C and D respectively, alongside two RNA markers (1 Kb:M_1 and 100 bp:M_2). The positions of the size markers are indicated in Kb.

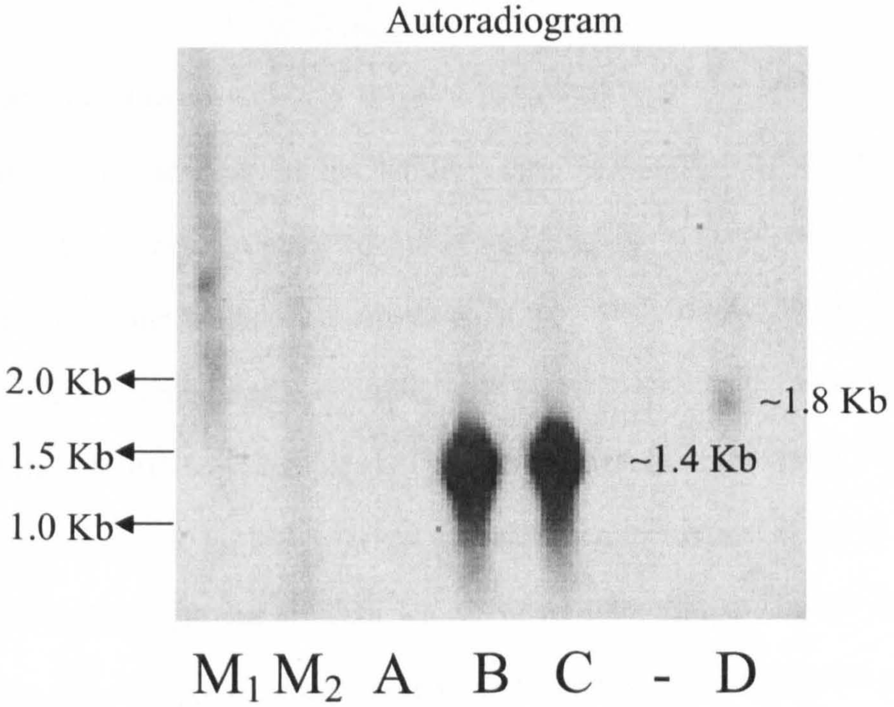
d) Northern blotting. Blotting was performed onto a Hybond+ membrane (see materials and methods) and the blot was hybridized to an anti-*gfp* probe and washed under stringent conditions (see materials and methods). Autoradiography was performed for 14 days before development (see materials and methods), and the autoradiogram is shown with the lanes indicated as well as the sizes of the RNA markers.


a)
Uninfected
Untransfected
Controls
pEF
pIR-1019
HOECHST
GFP-Expression
0.132
8.852e-
26.92
0.023
4.971
0.038
0.686
1.86
0.511
1.694
1.03
3.019
5.597
0.259
4.703
0.19
5.510
0.243
b)
Percentage GFP positive
0
10
20
30
40
50
60
70
80
90
iRBC1
iRBC2
pEF-1
PEF-2
pEF-3
piR1019-1
pIR1019-2
pIR1019-3
Experiment

c)


RNA gel
2.0 Kb
1.5 Kb
1.0 Kb
M1M2 A B C - D

d)


Autoradiogram
2.0 Kb
1.5 Kb
1.0 Kb
~1.8 Kb
~1.4 Kb
M1 M2 A B C - D

The GFP staining in the HOECHST *–ve* fraction (lower right quadrant, pEF Fig. 8.2 a) was probably due to GFP in cellular debris.

Transfection with the pIR1019 vector containing the M4 motif, however, resulted in low numbers of GFP positive iRBCs (Fig. 8.2 a and b). Approximately 4% of the pIR1019 transfected iRBCs were GFP positive, compared with 1% or less in the two untransfected iRBC controls (mock transfection controls). This showed that the construct containing M4, pIR1019, was only able to drive a very low expression in a few cells.

The level of transcription in the pEF and pIR1019 transfected iRBCs was also investigated. To do this, RNA was extracted (Fig. 8.2 c) and Northern blotting was performed (Fig. 8.2 d) with an anti-*gfp* probe (See also 2.10.2 to 2.10.3, Chapter II).

An overnight exposure did only revealed hybridisation of the probe to a 1.4 Kbp band in the lanes loaded with the two pEF, positive control samples. Only after 14 days (Fig. 8.2 d), could a weak band of around 1.8 Kbp be observed in the pIR1019 sample lane, while nothing was observed in the iRBC (mock transfection control RNA) lane confirming that hybridisation had occurred specifically to *gfp*. The larger size of the pIR1019 band compared to the two pEF bands, is consistent with the very long UTRs observed for the *yir* genes. The pEF promoter region in the pEF samples was 599 bp in length and it is not known where transcription initiates on the pEF promoter, but the approximately 400 bp size difference between the pEF-*gfp* and the pIR1019-*gfp* seem to suggest that for pIR1019-*gfp*, transcription had initiated within the IR1019 region. Thus the IR1019 region was able to drive a very low level of reporter gene transcription.

8.2.5 Transfection in *P.yoelii* YM (Transfection 2).

In order to test the effect of having removed M4, pIR894 was included in a second transfection experiment. In addition pTub (-) was also included as a negative control to assess if the very low level of GFP expression observed above was close to the background expression produced from a promoterless vector. In an attempt to obtain a higher percentage of transfected parasites, the lethal *P.yoelii* YM strain was used for transfection (See 2.9.11, Chapter II). This strain can invade normocytes, and thus infect a greater number of RBCs. This would hopefully increase the level of parasitaemia. Mice were injected with pIR1019, pIR894, pTub (-) and pEF transfected *P.yoelii* YM parasites along with untransfected and mock-transfected parasites. Parasite development in the infected animals was monitored and parasitaemia curves are shown (See supplementary, S8.4).

FACS analysis of the samples from this experiment (Fig. 8.3 a) was performed as described (See 2.10.1, Chapter II), after having bled the mice at the end of the third passage. The percentages of double HOECHST/GFP positive events were calculated (Fig. 8.3 b), and for the pEF transfected parasite lines, the percentage was around 70%. By contrast, transfection with pIR1019 or pIR894 only led to around 2% and 3% respectively. This was not different from the pTub (-) negative control.

The mean fluorescence intensity (MFI) was also calculated for both mononucleated and multinucleated iRBC (Fig. 8.3 b). A higher MFI was detected in all multinucleated cells, which is as expected as especially mature schizonts contains several merozoites, each capable of contributing to GFP expression.

Figure 8.3

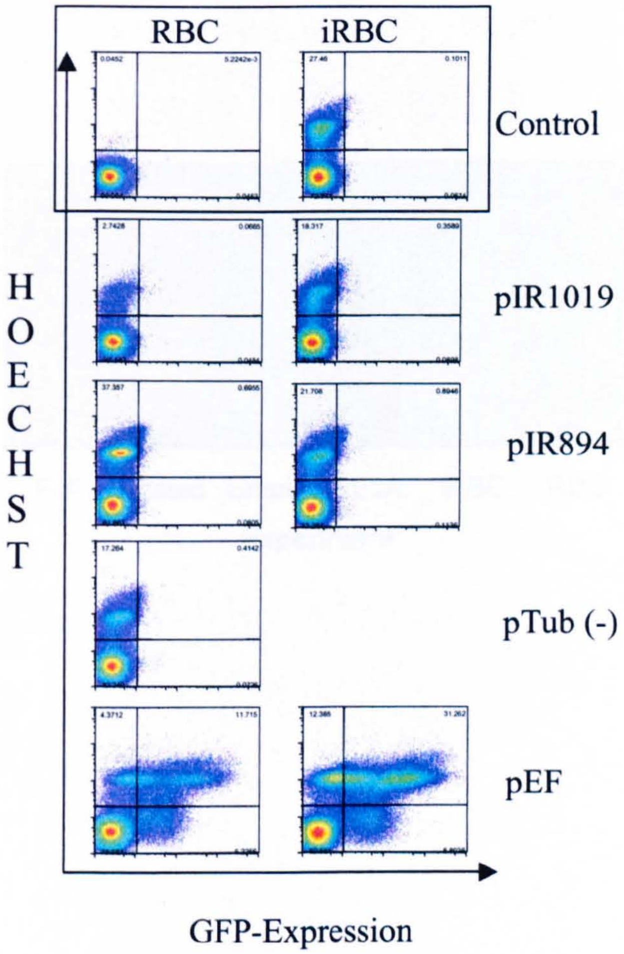
YM FACS analysis

a) FACS analysis of YM transfection experiment. The FACS plots (from top to bottom) shows: Controls (only controls contain both untransfected RBC and iRBC), pIR1019 transfected lines, pIR894 transfected lines, pTub (-) transfected line (negative control), pEF transfected lines (positive control). The X axis (FL1) shows the GPF expression and the Y axis (FL4) shows the HOECHST staining intensity. Each sample was measured in duplicates, and 50000 events were collected per sample.

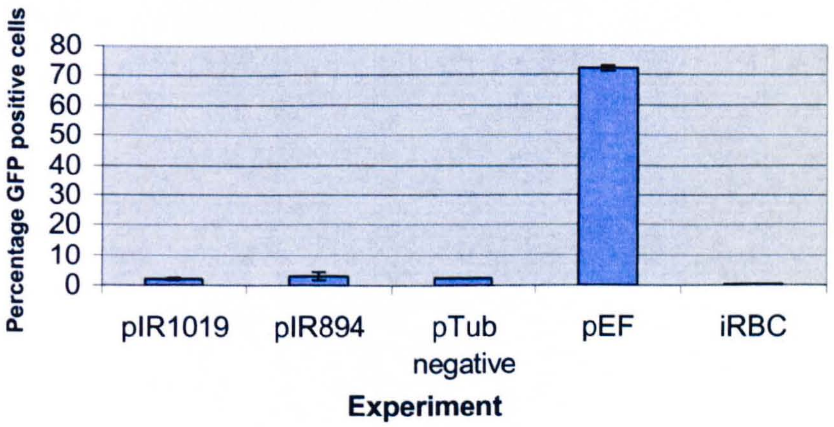
b) Percentage of double positive HOECHST/GFP events for all samples (from left to right (pIR1019, pIR894, pTub (-), pEF, iRBC). Where more than one line was transfected, the percentage is an average of these lines and standard deviations are based on these.

c) Mean fluorescence intensity (MFI) for all the samples. The events were divided into whether they corresponded to mono- or multi- nucleated cells (see materials and methods). The MFI for pIR1019, pIR894 and pTub negative, calculated separately for mono and multinucleated cells. Experimental standard deviations were included where more than one parasite line had been present.

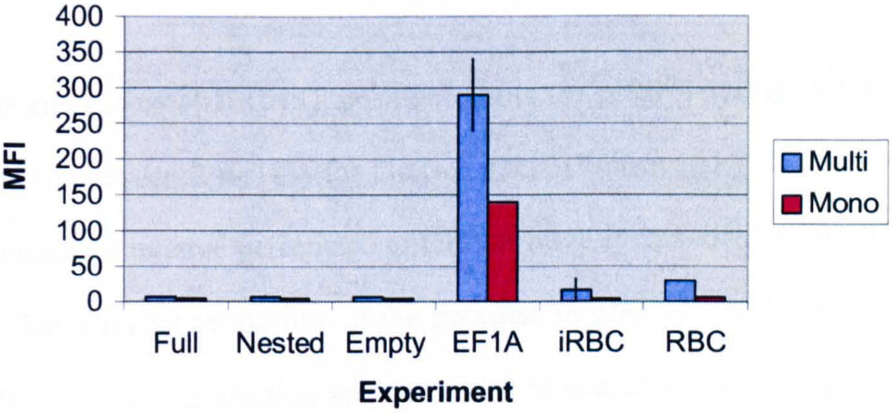
a)


RBC
iRBC
Control
pIR1019
pIR894
pTub (-)
pEF
HOECHST
GFP-Expression
0.0452
5.2242e-3
27.46
0.1011
2.7428
0.0665
18.317
0.3589
37.357
0.6956
21.706
0.6946
17.264
0.4142
4.3712
11.715
12.388
31.262

b)


Percentage GFP positive cells
80
70
60
50
40
30
20
10
0
pIR1019
pIR894
pTub negative
pEF
iRBC
Experiment

c)


400
350
300
250
200
150
100
50
0
MFI
Full
Nested
Empty
EF1A
iRBC
RBC
Experiment
Multi
Mono

The positive controls, pEF, exhibited a much higher MFI than pIR1019, pIR894 or pTub (-). In fact the three latter had similar levels of MFI. This showed that the transfection had worked, but the two experimental constructs (pIR1019 and pIR894) had comparable percentages and MFI to that of the negative control (pTub (-)).

High percentages of HOECHST positive events for pIR894-1 (Fig. 8.3 a, 21-37% compared to between 2 to 18% for the two pIR1019 lines and 17% for the pTub(-) line) reduced the relative percentage of GFP positive events for this line. This could indicate that a higher proportion of the parasites in pIR894-1 had lost the plasmid, especially since no transfection marker (which would allow estimating how large a proportion of the parasites still retained the plasmid) was present in this experiment. However, as all the mice were kept under drug cover until the end of the experiment, this is considered unlikely.

The mean fluorescence intensity (MFI) (Fig. 8.3 d) was thought to circumvent this problem, as this only measured the intensity of the GFP positive cells and would therefore allow a relative comparison of the effect of removing M4 and also for a comparison to the negative control. In conclusion, no differences in MFI between the two experimental constructs and the negative control could be established (Fig. 8.3 d). The fact that more GFP positive cells were seen in the pTub (-) transfected parasites than in the mock transfected probably reflect some leakiness from the pTub (-) construct. Unfortunately this leakiness was at a comparable level to the two experimental constructs.

8.3 Summary and discussion

In this Chapter, the role of the conserved upstream motif (M4) as a putative *cis* regulatory element was examined through transient transfections. However, incorporation of M4 into the vector in front of *gfp*,

did not result in good GFP expression. Thus, the transfection experiments did not clarify the role of M4. The major reasons for this outcome were thought to be:

1. The upstream region is too short. Although no conserved motifs were found upstream of M4 in the MEME analysis (Chapter VII), some possible enhancer regions could have been missed by delimiting the intergenic region to contain only M4. This is thought to be the most likely explanation.

2. The intergenic regions used did not constitute a functional *yir* promoter. The intergenic region, IR1019 (+M4), was not a complete match to known database sequences. Especially a few differences were observed (see position 265-280 in Fig. S8.1) in the region, where transcription was found to initiate (M2, see Chapter VII). However, this site was not universally conserved among *yir* genes (see Chapter VII), and it (M2) was not identified at the time these clones were constructed. Therefore it is possible that this particular intergenic region was not the strongest *yir* promoter.

3. GFP is not strong enough as a reporter molecule. GFP is the molecule of choice when it comes to live cell imaging and FACS analysis, however as its excitation and emission are discrete events, no signal accumulation occurs.

The Northern analysis shows that the pIR1019 (+M4) construct is not devoid of all ability to drive transcription of *gfp*. This was also confirmed by RT-PCR (not shown). However, the level of transcription was much lower than the positive control, which was also reflected in the FACS analysis, where virtually no GFP expression could be measured from pIR1019. Earlier studies of the *var* genes also showed a low level of expression compared to a housekeeping gene (Voss et al., 2000). From a large *P.vivax vir* EST analysis, a surprisingly low number of *vir* transcripts were detected (Merino et al., 2003). From this analysis of the *P.yoelii* EST database (at http://www.plasmodb.org), only a few *yir* transcripts has been detected. In addition, attempts in our laboratory of direct *yir* mRNA quantification (Northern blotting and Primer extension) have proven almost impossible. All this suggests that *yir* transcripts are of a very low abundance, and therefore a very low level of transcription would also be expected. However, from the FACS analysis, expressions from the pIR1019 (+M4) and pIR894 (-M4) constructs were not different from the negative control without any promoter, pTub (-). This is probably caused by any of the reasons mentioned above, although it cannot be excluded that a *yir* promoter has a very low intrinsic activity. In further experiments, the three possible causes: 1: Not a functional promoter, 2: Too short 5′ intergenic region and 3: GFP too weak, would have to be addressed. In order to make sure, the promoter is functional; a couple of intergenic regions from the highly abundant transcripts (see Chapter III) should be used, while increasing the upstream length of these. To increase sensitivity, *luciferase* could be used as a reporter gene instead of *gfp*. Alternatively, yir promoters could be fused to a drug resistance gene as performed in a recent *var* gene study (Gannoun-Zaki et al., 2005). If a *yir* promoter is only active in a small number of cells, this would allow for selection of those.

Chapter IX

Alternative splicing

9.1 Introduction

Alternative splicing has been suggested to be another important level of regulation in *Plasmodium*. In a comparative genomic analysis between *P. falciparum* and other unicellular eukaryotes, it has been found that *P.falciparum* contained a relatively lower level of identifiable transcription associated factors (TAF), but a higher level of mRNA processing factors (Coulson et al., 2004), and in two studies combining microarray analysis with proteome analysis, several transcripts appeared earlier than the protein in *P. falciparum* and *P. berghei.* (Fan et al., 2004 and Hall et al., 2005). Stage specific splicing patterns has been found in *P.yoelii* and *P.falciparum* for *maebl* (Singh et al., 2004) and *stevor* transcripts (Sutherland et al., 2001), and it is therefore likely that mRNA splicing is an important level of regulation of protein expression.

9.1.1 Objectives

One unexpected major *yir* UTR splice form, along with two minor splice variants were identified (9.2.1) along with unspliced *yir* UTRs (9.2.2). All the introns of the major form were highly conserved and contained complementary flanking regions (9.2.3), and these introns were distinctly distributed among the *yir* supergroups (9.2.4). In an attempt elucidate what the function(s) of these splicing events could be, it was investigated if translatable in-frame alternative first exons could be produced from one of these splicing types (9.2.5). RNA structure predictions were performed to see if a distinct structure resulted from the splicing (9.2.6). It was also investigated if the splicing removed upstream open reading frames (uORF) in the *yir* intergenic regions (9.2.7).. Finally, the spliced 5′intergenic regions were scanned for their presence of putative splicing factor binding sites (9.2.8), and the mutually exclusive, stage specific, BIR expression pattern (Hall et al., 2005) were analysed in the light of this splicing pattern (9.2.9).

9.2 Results

9.2.1 Characterisation of three splice variants

During 5′ RACE and in RT-PCR reactions (see Chapter III and VII), alternative splicing was found in the 5′ intergenic regions. Five of the six sequenced 5′ RACE transcripts were found to be spliced in a similar region some 120 nt (position of intron acceptor site) upstream of the *yir* exon 1. This splice variant, called Sv1, led to the removal of introns with a length of approximately 110 to 120 nt. All the introns started with GT and ended with AG, and an alignment between one of the sequenced transcripts and it's corresponding contig can be seen in the supplementary section (See supplementary, S9.1).

To investigate this splicing pattern in more detail, and also to see if it changed in the different blood stages of *P.yoelii*, a primer set located just before the UTR intron and in exon 2 (Fig. 9.1 a) was designed and RT-PCR was performed (Fig. 9.1 b). For Schizont-Trophozoite- and -Ring stage parasites, two bands could be seen at around 590 and 490 nt as compared to the single 800 nt band seen from a PCR with the same primer set on genomic DNA (Fig. 9.1 b). The 590 nt band corresponded roughly to the sizes of the expected product if both the UTR intron (around 110-120 nt) and the 1st intron within the gene (around 108-120 nt) were removed.

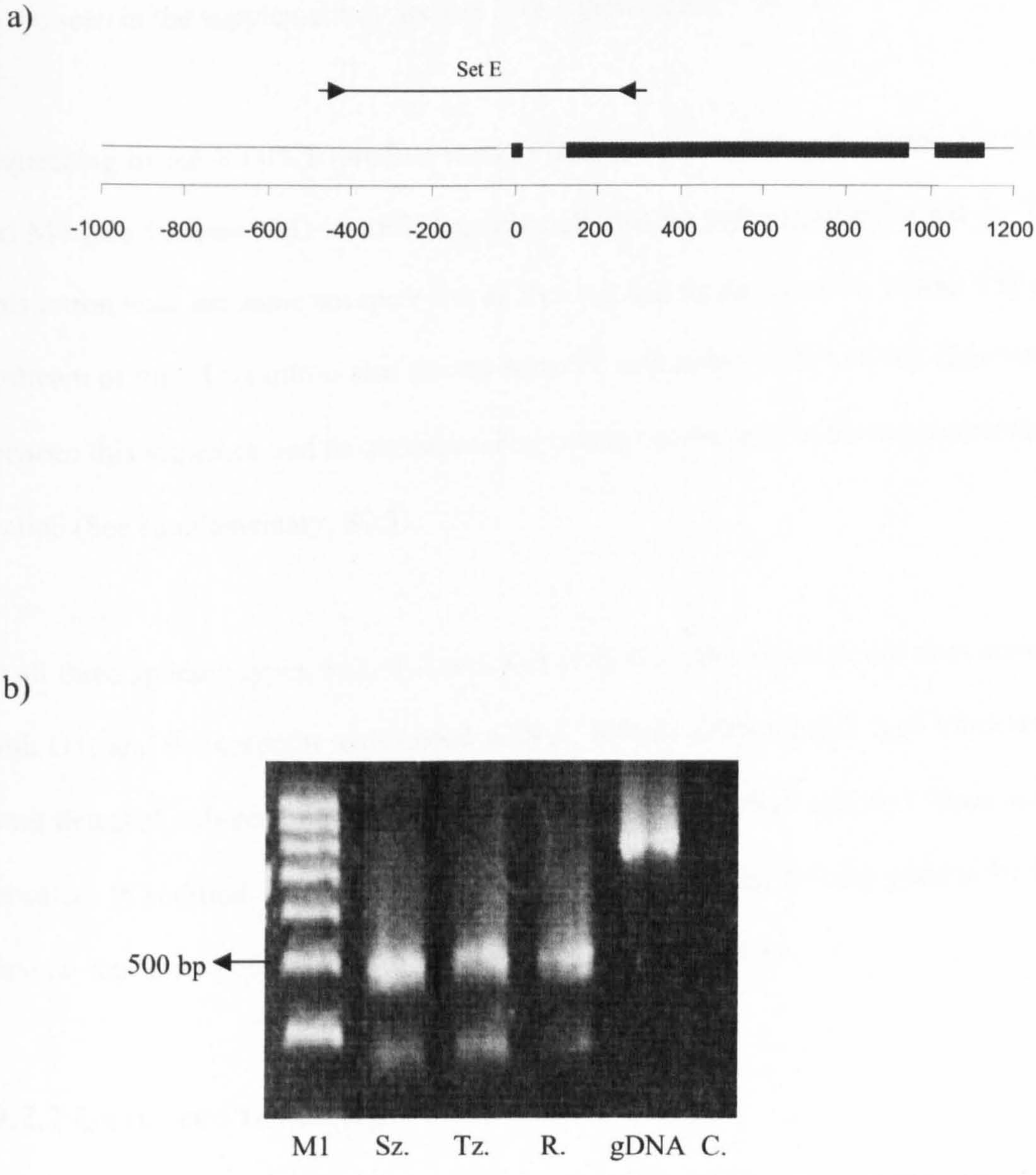
Sequencing of the RT-PCR products from the isolated 590 nt fragment, identified 8 different transcripts all of which were Sv 1 types of splicing. The 490 nt fragment was also isolated and sequenced, and this identified an additional splice variant

Figure 9.1.

Investigation of alternative splicing in three blood stages

a) Location of the used primer pair, set E The expected size range of the products amplified fragments using this set was: 700-800 (gDNA), 575-675 (cDNA, intron 1 only) and 460-560 (cDNA, both intron 1 and UTR intron). The forward primer was localized immediately before the donor site of the UTR intron.

b) PCR was performed on cDNA from stage-separated parasites from Schizont (Sz), Trophozoite (Tz) and Ring (R) stages. PCR was performed simultaneously on both cDNA from the three stages and gDNA as well as a negative control sample. In all the three cDNA samples, two bands of 490 and 390 nt could be observed, while only one band of around 800 nt can be seen from the PCR on genomic DNA.


a)
Set E
-1000
-800
-600
-400
-200
0
200
400
600
800
1000
1200
b)
500 bp
M1
Sz.
Tz.
R.
gDNA
C.

(called Sv 2), where the donor site of the UTR intron had been joined with the acceptor site of the 2nd exon. This intron was 373 nt long and started with GT and ended with AG. An alignment between this sequence and its corresponding contig can be seen in the supplementary section (See supplementary S9.2).

Sequencing of the RT-PCR product used to map the transcription initiation site M2 and M4 (see Chapter VII) identified another alternative splicing type (called Sv 3). This intron used the same acceptor site as Sv1 but had its donor site a further 453 nt upstream of this. This intron also started with GT and ended with AG. An alignment between this sequence and its corresponding contig can be seen in the supplementary section (See supplementary, S9.3).

In all three splicing types Sv1, Sv2 and Sv3 (Fig. 9.2), the intron donor sites started with GT, and the acceptor sites ended with A. Fifteen different Sv1 type transcripts were detected, whereas only one transcript of each of the Sv2 and Sv3 types were detected. In addition, for all the Sv1 and Sv3 types, a normal splicing pattern for the first *yir* intron was observed (See supplementary, S9.1 to S9.3).

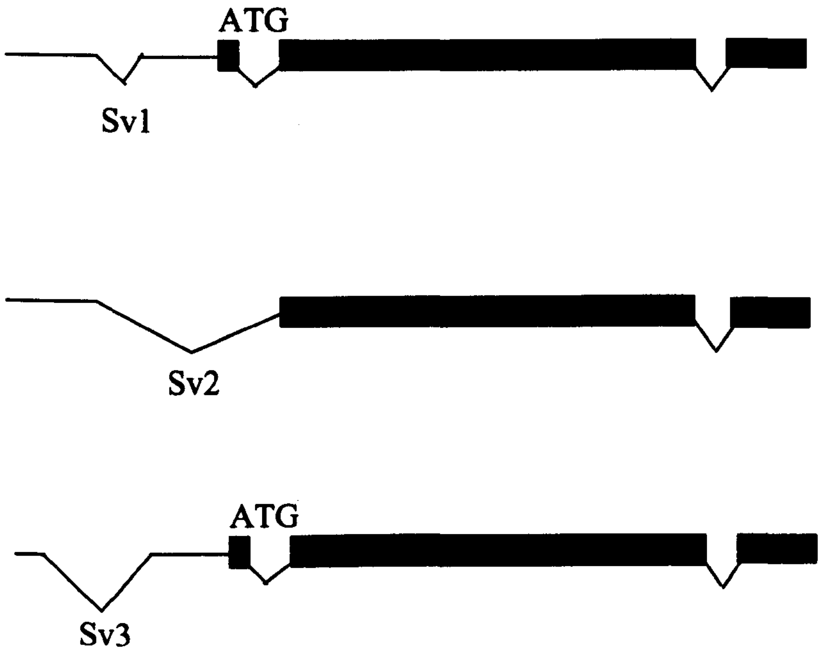
9.2.2 Unspliced transcripts

One unspliced UTR transcript was detected by 5′ RACE and another unspliced transcript was found in the rodent malaria EST database at www.plasmodb.org. Both of these were spliced between the first and second exon, indicating that they are both from processed mRNA. In these two UTRs, no regions resembling the typical UTR intron flanking or the intron itself were identified. This showed that some types of transcripts contained the intron in their UTRs while others did not. In the

Figure 9. 2

Schematic of the three alternative splicing variants (Sv 1, 2 and 3)

Sv 1 was the most commonly found and was found in a total of 16 *yir* transcripts. Sv 1 splicing occurred in a specific region of the 5′ UTR around 110 nt upstream of the ATG, and lead to the splicing out of introns of around 110 nt. Sv 2 used the same donor site as Sv 1, but bypassed the acceptor site, used by Sv 1 splicing types, and instead used the start of the second exon as an acceptor site, leading to skipping of exon 1. Sv 3 used a donor site located 430 nt upstream of the ATG and an acceptor site close to the acceptor sites used for Sv 1 splicing types.


ATG
Sv1
Sv2
ATG
Sv3

supplementary section, the alignments between these sequences and their respective contigs can be seen (See supplementary, S9.4).

9.2.3 Regions flanking Sv1

To determine whether differences in flanking regions or intron structure could explain the alternative splicing observed, these regions of the Sv 1 type were aligned (Fig. 9.3). The regions flanking the introns were highly conserved, and from position 10 to 15 before the donor site, the consensus sequence *AACCCT* occurred for all but three of the sequences. This was followed by another complementary consensus sequence located after the acceptor sites for the majority of the introns at position 154 to 158 with the consensus *AGGGTT*. Four transcripts either did not contain this consensus or had the donor site located in the *AG* of this consensus.

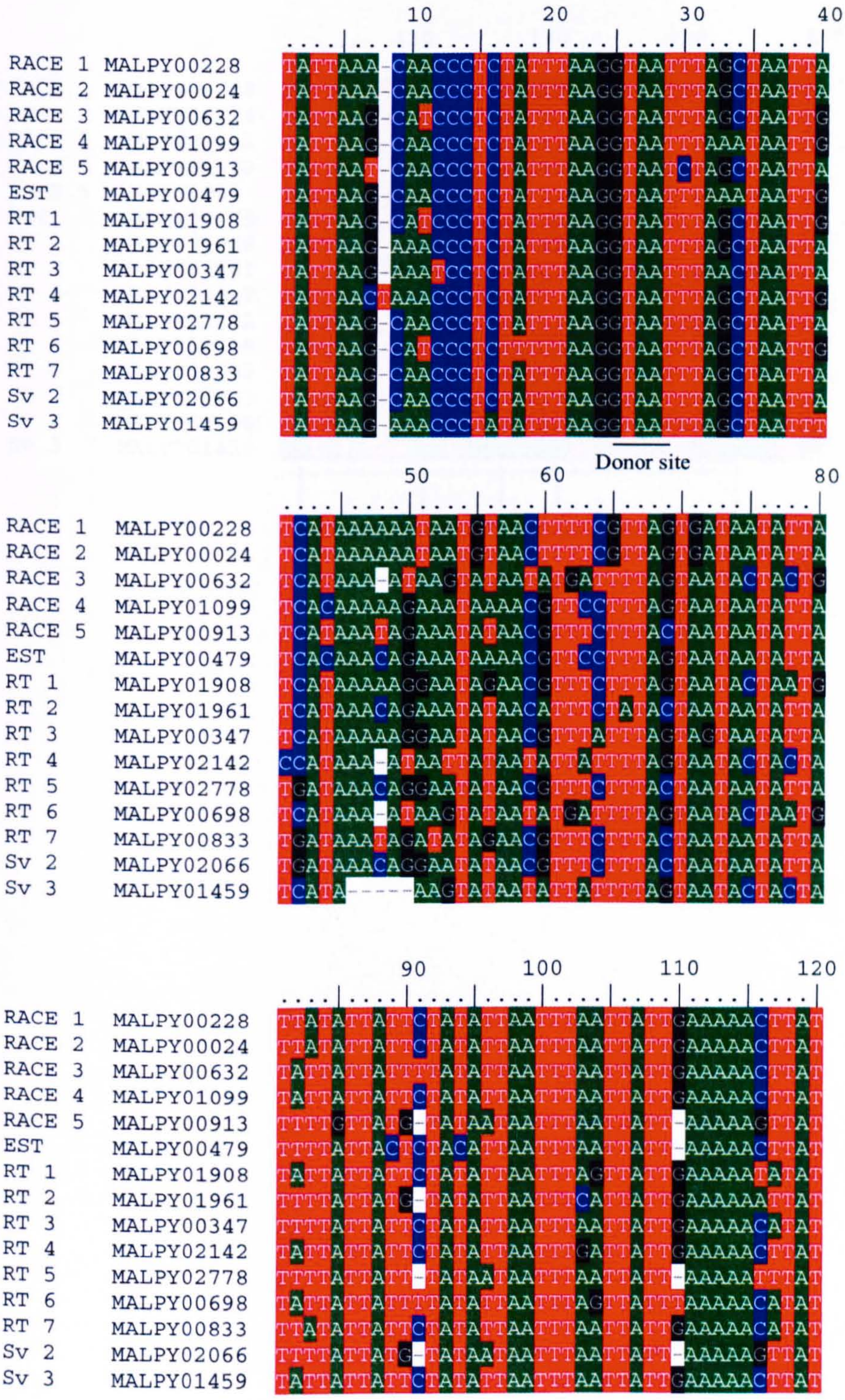
9.2.4 Yir supergroups and UTR splicing

As both spliced and unspliced UTR transcripts had been detected, it was important to investigate how these were distributed among the five *yir* supergroups (SG1-SG5, Chapter II), as this splicing could potentially have a regulatory function. This revealed (Fig. 9.4 a), that all the UTR spliced transcripts, except one, belonged to SG1. The two unspliced UTR transcript (see section 9.2.2) belonged to SG1 and SG3 respectively. The reason for detection of only one transcript outside SG1 could be explained by primer bias for the RT-PCR since the primer in the UTR was designed from a collection of UTRs spliced transcripts. For the 5′ RACE a universal exon 3 primer was used, and this region was highly conserved in genes from all the supergroups. However, 5 out of 6 transcripts from this belonged to SG1.

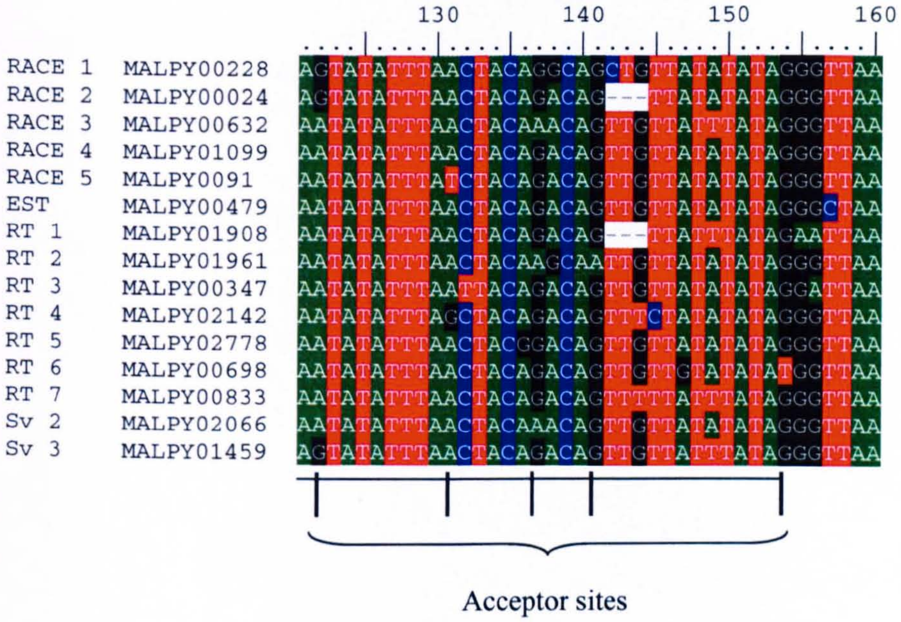
Figure 9.3

Introns and flanking regions for Sv 1 splice types

An alignment between the UTR spliced flanking regions and introns. Introns were deduced from alignment to contigs in each case. Underlined are the universally used donor site and the acceptor sites. Acceptor sites used for the individual transcripts were located at various positions within a stretch of 39 nucleotides. In most cases the acceptor site used was the first available "ag" within this stretch. The regions flanking the introns were highly conserved, and from position 10 to 15 before the donor site, the consensus sequence *AACCCT* occurred for all but three of the sequences. This was followed by another complementary consensus sequence located after the acceptor sites for the majority of the introns at position 154 to 158 with the consensus *AGGGTT.* Three of the spliced transcripts deviated somewhat with respect to the *GGG* nucleotides in the complementary region.



```
                                 10        20        30        40
                        ....|....|....|....|....|....|....|....|
RACE 1 MALPY00228       TATTAAA-CAACCCTCTATTTAAGGTAATTTAGCTAATTA
RACE 2 MALPY00024       TATTAAA-CAACCCTCTATTTAAGGTAATTTAGCTAATTA
RACE 3 MALPY00632       TATTAAG-CATCCCTCTATTTAAGGTAATTTAGCTAATTG
RACE 4 MALPY01099       TATTAAG-CAACCCTCTATTTAAGGTAATTTAAATAATTG
RACE 5 MALPY00913       TATTAAT-CAACCCTCTATTTAAGGTAATCTAGCTAATTA
EST    MALPY00479       TATTAAG-CAACCCTCTATTTAAGGTAATTTAAATAATTG
RT 1   MALPY01908       TATTAAG-CATCCCTCTATTTAAGGTAATTTAGCTAATTG
RT 2   MALPY01961       TATTAAG-AAACCCTCTATTTAAGGTAATTTAGCTAATTA
RT 3   MALPY00347       TATTAAG-AAATCCTCTATTTAAGGTAATTTAACTAATTA
RT 4   MALPY02142       TATTAACTAAACCCTCTATTTAAGGTAATTTAGCTAATTG
RT 5   MALPY02778       TATTAAG-CAACCCTCTATTTAAGGTAATTTAGCTAATTA
RT 6   MALPY00698       TATTAAG-CATCCCTCTTTTTAAGGTAATTTAGCTAATTG
RT 7   MALPY00833       TATTAAG-CAACCCTCTATTTAAGGTAATTTAGCTAATTA
Sv 2   MALPY02066       TATTAAG-CAACCCTCTATTTAAGGTAATTTAGCTAATTA
Sv 3   MALPY01459       TATTAAG-AAACCCTATATTTAAGGTAATTTAGCTAATTT
                                                 ____
                                             Donor site
                                 50        60                  80
                        ....|....|....|....|....|....|....|....|
RACE 1 MALPY00228       TCATAAAAAATAATGTAACTTTTCGTTAGTGATAATATTA
RACE 2 MALPY00024       TCATAAAAAATAATGTAACTTTTCGTTAGTGATAATATTA
RACE 3 MALPY00632       TCATAAA-ATAAGTATAATATGATTTTAGTAATACTACTG
RACE 4 MALPY01099       TCACAAAAAGAAATAAAACGTTCCTTTAGTAATAATATTA
RACE 5 MALPY00913       TCATAAATAGAAATATAACGTTTCTTTACTAATAATATTA
EST    MALPY00479       TCACAAACAGAAATAAAACGTTCCTTTAGTAATAATATTA
RT 1   MALPY01908       TCATAAAAAGGAATAGAACGTTTCTTTAGTAATACTAATG
RT 2   MALPY01961       TCATAAACAGAAATATAACATTTCTATACTAATAATATTA
RT 3   MALPY00347       TCATAAAAAGGAATATAACGTTTATTTAGTAGTAATATTA
RT 4   MALPY02142       CCATAAA-ATAATTATAATATTATTTTAGTAATACTACTA
RT 5   MALPY02778       TGATAAACAGGAATATAACGTTTCTTTACTAATAATATTA
RT 6   MALPY00698       TCATAAA-ATAAGTATAATATGATTTTAGTAATACTAATG
RT 7   MALPY00833       TGATAAATAGATATAGAACGTTTCTTTACTAATAATATTA
Sv 2   MALPY02066       TGATAAACAGGAATATAACGTTTCTTTACTAATAATATTA
Sv 3   MALPY01459       TCATA-----AAGTATAATATTATTTTAGTAATACTACTA

                                 90       100       110       120
                        ....|....|....|....|....|....|....|....|
RACE 1 MALPY00228       TTATATTATTCTATATTAATTTAATTATTGAAAAACTTAT
RACE 2 MALPY00024       TTATATTATTCTATATTAATTTAATTATTGAAAAACTTAT
RACE 3 MALPY00632       TATTATTATTTTATATTAATTTAATTATTGAAAAACTTAT
RACE 4 MALPY01099       TATTATTATTCTATATTAATTTAATTATTGAAAAACTTAT
RACE 5 MALPY00913       TTTTGTTATG-TATAATAATTTAATTATT-AAAAAGTTAT
EST    MALPY00479       TTTTATTACTCTACATTAATTTAATTATT-AAAAACTTAT
RT 1   MALPY01908       TATTATTATTCTATATTAATTTAGTTATTGAAAAATATAT
RT 2   MALPY01961       TTTTATTATG-TATATTAATTTCATTATTGAAAAAATTAT
RT 3   MALPY00347       TTTTATTATTCTATATTAATTTAATTATTGAAAAACATAT
RT 4   MALPY02142       TATTATTATTCTATATTAATTTGATTATTGAAAAACTTAT
RT 5   MALPY02778       TTTTATTATT-TATAATAATTTAATTATT-AAAAATTTAT
RT 6   MALPY00698       TATTATTATTTTATATTAATTTAGTTATTTAAAAACATAT
RT 7   MALPY00833       TTATATTATTCTATATTAATTTAATTATTGAAAAACATAT
Sv 2   MALPY02066       TTTTATTATG-TATAATAATTTAATTATT-AAAAAGTTAT
Sv 3   MALPY01459       TATTATTATTCTATATTAATTTAATTATTGAAAAACTTAT
```



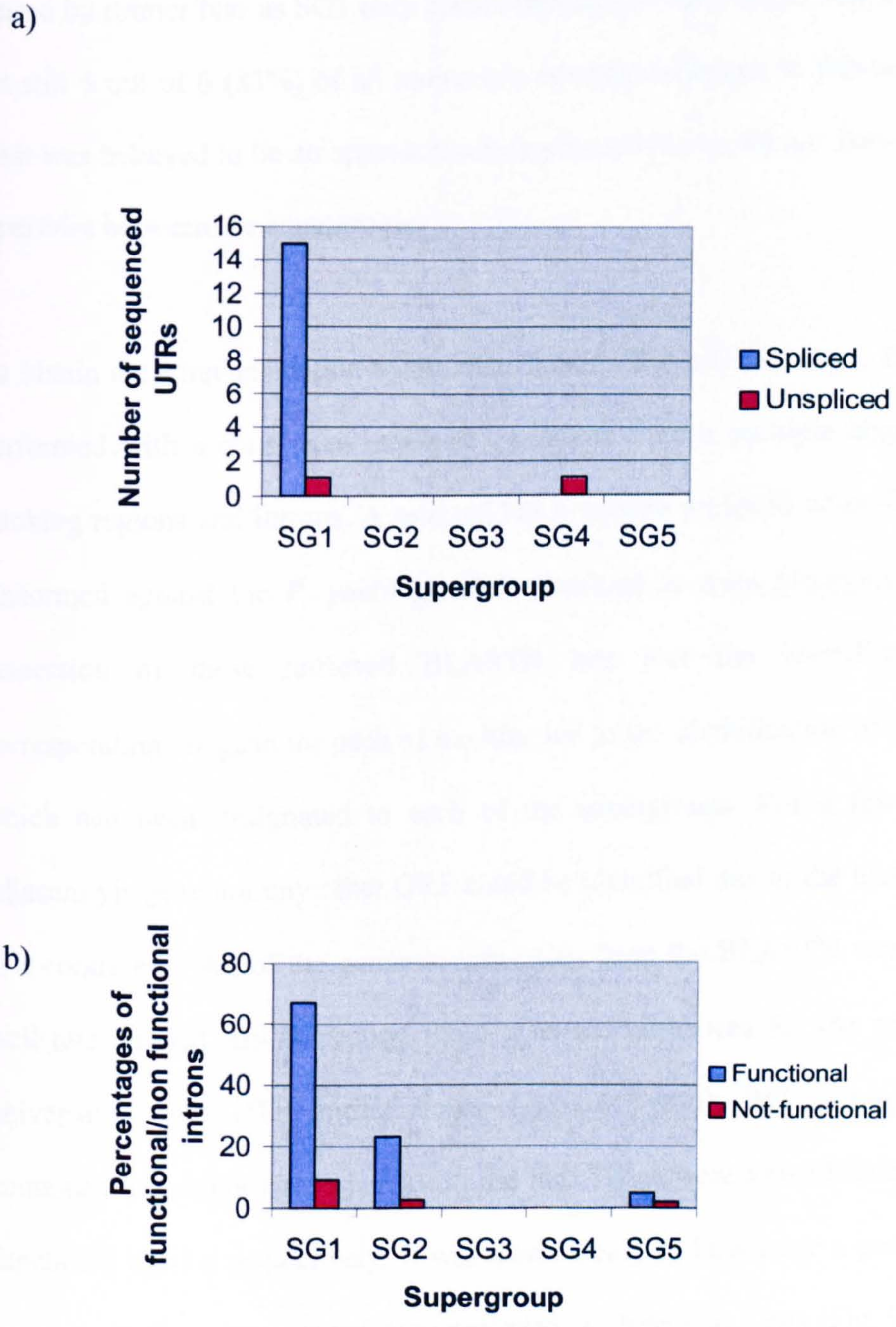
```
                                     130       140       150       160
                                ....|....|....|....|....|....|....|....|
RACE 1   MALPY00228             AGTATATTTAACTACAGGCAGCTGTTATATATAGGGTTAA
RACE 2   MALPY00024             AGTATATTTAACTACAGACAG---TTATATATAGGGTTAA
RACE 3   MALPY00632             AATATATTTAACTACAAACAGTTGTTATTTATAGGGTTAA
RACE 4   MALPY01099             AATATATTTAACTACAGACAGTTGTTATATATAGGGTTAA
RACE 5   MALPY0091              AATATATTTATCTACAGACAGTTGTTATATATAGGGTTAA
EST      MALPY00479             AATATATTTAACTACAGACAGTTGTTATATATAGGGCTAA
RT 1     MALPY01908             AATATATTTAACTACAGACAG---TTATTTATAGAATTAA
RT 2     MALPY01961             AATATATTTAACTACAAGCAATTGTTATATATAGGGTTAA
RT 3     MALPY00347             AATATATTTAATTACAGACAGTTGTTATATATAGGATTAA
RT 4     MALPY02142             AATATATTTAGCTACAGACAGTTTCTATATATAGGGTTAA
RT 5     MALPY02778             AATATATTTAACTACGGACAGTTGTTATATATAGGGTTAA
RT 6     MALPY00698             AATATATTTAACTACAGACAGTTGTTGTATATATGGTTAA
RT 7     MALPY00833             AATATATTTAACTACAGACAGTTTTTATTTATAGGGTTAA
Sv 2     MALPY02066             AATATATTTAACTACAAACAGTTGTTATATATAGGGTTAA
Sv 3     MALPY01459             AGTATATTTAACTACAGACAGTTGTTATTTATAGGGTTAA
```

Figure 9.4

Supergroup distribution of splicing

a) Shows the supergroup distribution of sequences, which were spliced in their UTR or not. A total of 15 sequences were found to be spliced and these belonged to SG1. One sequence, also belonging to SG1 was identified in the *P.yoelii* EST database at www.http://plasmodb.org was not spliced in the UTR, and likewise sequencing revealed a transcript in SG4 which were also not spliced. Neither of the two unspliced sequences contained the intron as identified by their lack of GT-AG sequences at donor and acceptor sites respectively.

b) A BLASTN (www.http://plasmodb.org) was performed with a consensus sequence created from the experimentally verified UTR Sv 1 type flanking regions and introns. All the hits were retrieved and aligned. By visual inspection, it was determined how many contained functional donor and acceptor sites, and the supergroup identity of each of the hits were determined. The graph shows the percentages of genes in each supergroup, which were classified as containing either, a functional or non-functional UTR intron.


a)
Number of sequenced UTRs
16
14
12
10
8
6
4
2
0
SG1
SG2
SG3
SG4
SG5
Supergroup
Spliced
Unspliced
b)
Percentages of functional/non functional introns
80
60
40
20
0
SG1
SG2
SG3
SG4
SG5
Supergroup
Functional
Not-functional

The relative proportion of transcripts from other supergroups compared to the number of SG1 transcripts in the 5′ RACE seemed to indicate that this was not just caused by primer bias as SG1 only constituted around 40% of the total *yir* repertoire, but still 5 out of 6 (83%) of all transcripts detected belonged to this supergroup, in what was believed to be an approach where primer bias would not skew the detected repertoire between the supergroups.

To obtain a clearer indication of the distribution of this UTR intron, BLASTN was performed with a consensus sequence generated from a multiple alignment of the flanking regions and introns. A total of 314 hits were retrieved when BLASTN was performed against the *P. yoelii* genomic database at www.plasmodb.org. Manual inspection of these retrieved BLASTN hits and the identification of the corresponding *yir* gene for each of the hits, led to the identification of 290 *yir* genes, which had been designated to each of the supergroups. For a few hits, neither adjacent yir gene nor any other ORF could be identified due to the localisation close to a contig end. All of the genomic sequences from the BLASTN were retrieved as well and aligned. By inspecting these genomic sequences for the presence of the universally used UTR intron donor site (*AAGgtaa*), 255 genomic sequences contained this donor site, whereas 35 did not. These were termed functional or non-functional introns respectively. It was then calculated how large a proportion of the genes in the five supergroups were preceded by these two types (Fig. 9.4 b). In total, 67% of SG1, 23% of SG2 and 4.8% of SG5 contained putative functional UTR introns. Also, manual inspection of 306 5′ intergenic regions (see Chapter VII) from all the supergroups, led to the identification of 128 Sv1 type introns. Of these, 117 out of the 146 (80%) SG1 5′intergenic regions contained an intron with the conserved consensus (*AAGgtaa*). The remaining 20% were *AAGgcaa, AAGgtag,* *AAGatag* or *AAAgtaa* in the conserved donor site. The *AAGgtag* and *AAAgtaa* types started with GT and could potentially be spliced, although none of these types were seen experimentally. The remaining 11 Sv1 type introns were located in SG2 or SG5. The estimate is therefore that 67-80% of SG1 genes contained the Sv1 type intron in their UTRs, while 23% of SG2 and 4.8% of SG5 did so. Since SG2 to SG5 were all supported by high bootstrap values (See Chapter IV), the location of introns in the 5' UTR of genes from these supergroups, most likely reflect some mixing of function attributes between the supergroups. This clearly showed that the UTR intron is highly present in the UTR of SG1 genes and a proportion of SG2 genes also contain introns in their UTRs. As SG1 and SG2 were found to be the supergroups with the least proportion of their genes located on the subtelomeric contigs (see Chapter IV), it is interesting to note that the distribution of this UTR intron follows this quite clearly, with none of SG3 or SG4 genes and only a very low level (4.8%) of SG5 genes being found to contain this intron. This showed that the UTR intron is distinctly distributed among the supergroups

9.2.5 Alternative first exons

Sv 1 and Sv 3 did not alter the coding potential of the mRNA as the first coding *yir* exon was unaffected by the UTR splicing. In contrast to this, Sv 2 did have an effect on the coding potential through the removal of exon 1 from the mRNA. This mechanism could have brought together a variant coding exon 1 located upstream of the UTR donor site with that of the coding *yir* exon 2, however no such upstream coding exon could be identified. It though still remained a possibility that such a coding exon was present in other transcripts also containing this type of UTR intron. By manual inspection of the annotated genes, two other *yir* genes: PY01578 and PY04964 were identified to contain such an alternative coding exon 1 in frame with the second *yir* exon. For these two genes, the automated gene annotation performed by TIGR had predicted the exon 1 skipping event, Sv 2, and removed an intron similar to the one observed experimentally. Since this splicing was found to occur for one transcript, there would therefore be a possibility that it could happen for other transcripts as well.

To analyse how widespread alternative coding exons were, a BLASTN using a 60 nt region immediately before the UTR donor site in Sv 2, was performed. This retrieved 248 hits, and from these, three types of in-frame alternative coding exons (called alternative exons, AE 1 to 3) were identified in a total of 19 sequences. These were identified for 19 *yir* genes out of 248 analysed (7.6%), while the remaining 227 did not contain these alternative-coding exons. Three types of alternative exons, named AE1 to AE3 were identified based on coding sequence length. When these were translated into amino acids (Fig. 9.5), it can be seen that these three types were highly similar but differed progressively in size, as described above, from 20 to 9 amino acids. The alternative exons identified here were further analysed for targeting signals to the apicoplast or mitochondrion at www.plasmodb.org, for signal or cleavage peptide similarity at http://www.cbs.dtu.dk and for Pfam domains at http://www.sanger.ac.uk/. These sequences did not result in any BLASTP hits.

A BLASTP search against all databases at http://www.ncbi.nlm.nih.gov retrieved only hits in *Plasmodium yoelii*. This showed that these putative peptides were only present in *Plasmodium yoelii*, and it could not be determined if they were involved in transport or cleavage processes by comparing to known peptides involved in these events.

Figure 9.5

Alternative first exons

Type A, B and C alternative exons shown as amino acid sequences. These 19 sequences were the only ones that could be identified by BLASTN analysis to start with ATG, and be in-frame with the *yir* second exon if spliced by the Sv 2 type alternative splicing.

Alternative exon type A.

```
                                10        20
                           ....|....|....|....|
chrPyl_00423|MALPY         MQNYIIYTYDKRCIKRPQNK
chrPyl_01554|MALPY         MKNYIIYTYGKRCIKRPPNK
chrPyl_00092|MALPY         MQNCIIYTYDKTCIKHRPNK
chrPyl_01902|MALPY         MQNCIIYIYDKKCLNTPQNK
chrPyl_00691|MALPY         MQNCIIYAYDNKCLNVPQNK
chrPyl_02253|MALPY         MQNCIIYTYDKKCINALQNK
chrPyl_01574|MALPY         MQNCIIYTYDKTYIKSTQNK
chrPyl_01417|MALPY         MQNCIIYTYDNKCLNTPQNK
```

Alternative exon. Type B.

```
                               5    10
                           ....|....|..
chrPyl_01110|MALPY         MIKHVLNIPQNK
chrPyl_00393|MALPY         MIKHVLNISQNK
chrPyl_01302|MALPY         MIKYVLNNLPNK
chrPyl_01875|MALPY         MIKNVLNNPQNK
chrPyl_02765|MALPY         MIINVLNPPQNK
chrPyl_02058|MALPY         MIKNVLNTPQNK
chrPyl_00431|MALPY         MIINVLNIPQNK
chrPyl_00843|MALPY         MIKHVLNIPQNK
chrPyl_00484|MALPY         MIKNVLNTPQNK
```

Alternative exon type C.

```
                               5
                           ....|....
chrPyl_02382|MALPY         MFKRPPQNK
chrPyl_02737|MALPY         MFKHPQNK
```

9.2.6 RNA structure predictions

The two complementary regions could play a role in folding the RNA molecule in such a way that this would promote splicing, or, they could have an effect on the mRNA structure after the intron was removed. This was investigated by aligning 110 *yir* UTRs, found to contain the Sv1 type intron (Fig. 9.6). In this alignment, the structure was predicted both with the intron present and with the intron removed. Structures of these two alignments were predicted at: www.http://www.genebee.msu.su/services/rrna2_reduced.html.

The predicted mRNA structure for sequences both with (Fig. 9.6 a and b) and without (Fig. 9.6 c and d) the intron was analysed. The RNA structure with the intron did not use the two complementary regions to form a stem, but instead a stem was formed by interactions between a region located just before the *AA(CCC)T* consensus and a region close to the donor site of the intron (Fig. 9.6 b). The RNA structure without the intron formed a stem by interactions between a region just before the *AA(CCC)T* consensus (not the same as Fig. 9.6 b) and a region immediately after the region flanking the introns acceptor site (Fig. 9.6 c). In this case, the RNA molecule assumed the "clover leaf" or "Y" shaped structure, which characterises IRES sequences (Pesole et al., 2001, see introduction).

9.2.7 Open reading frames in the untranslated regions

Open reading frames (ORFs) located in untranslated regions (uORFs) have often been found to decrease the amount of translation that can occur from an mRNA molecule (see introduction, Chapter I). One possible role for the UTR splicing could be to decrease the number of ORFs and thereby increase the probability of translation of certain transcripts. To investigate this, 306 5′ intergenic regions (see Chapter VII)

Figure 9.6

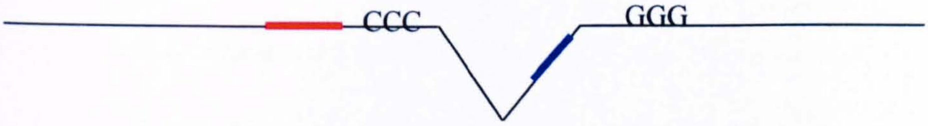
RNA structure predictions

Multiple alignments were made of 110 *yir* 5′ intergenic regions ranging from M2 (approximately 810 upstream of the ATG) to immediately before the ATG. In one alignment, the UTR intron (Sv1 type) was include, whereas it was removed in the other. RNA structures were predicted at http://www.genebee.msu.su/services/rna2_reduced.html (see materials and methods). Apart from the predicted structures, the program gives an overview over the positions of predicted stems and their energy. This was used to investigate which regions in the vicinity of the intron before and after splicing participated in stem formations.

a) A schematic of identified stems. The two complementary flanking regions are shown in letters (CCC and GGG) and the intron. In red and blue is shown the regions that participated in a stem. Below: Sequences of the red and blue regions are shown in order of appearances in bold. Intron sequence is underlined and splice site junctions are shown in underlined italics.

b) RNA structure for 110 5′ intergenic regions containing the intron. On the figure is indicated which of the stems involved regions in the vicinity of the intron. As can be seen, a region before the CCC regions and a region towards the donor site within the intron were formed a stem with a free energy of –10.7 Kcal/mol, which was the second lowest free energy in the entire structure.

a)



ATTATTTTTATTTTTTAGATTTATATAGTACTTAAGTTTTAGAATTGAAA
TCACTAAAACAGAAGATATAAATATATTCCTTTTCTATGTTCAAAAATAC
GTTAAATTCATTATAAAAGTATATCCATTTTTATAATAAAGTTATCCAGAT
GCTAGTAAGAATAGCATTTTTGTATAAAATGCATCATATCTATATTTAAT
CAA**AATGTATTA**AGCAACCCTCTATTT*AAG*GTAATTTAGCTAATTATGA
TAAACAGGAATATAACGTTTCTTTACTAATAATATTATTTTATTATATATA
ATAATTTAATTATTAAAAAGTTA**TAATACATTT**AACTACAG*ACAG*TTGTT
ATATATAGGGTTAAAGTATTTGCGTCACATATTGGGGACAACGTATATAA
AGCAGCATGATTCTACTCAATTTACAAATTAAAAATAAAATCCCATTATA

b)

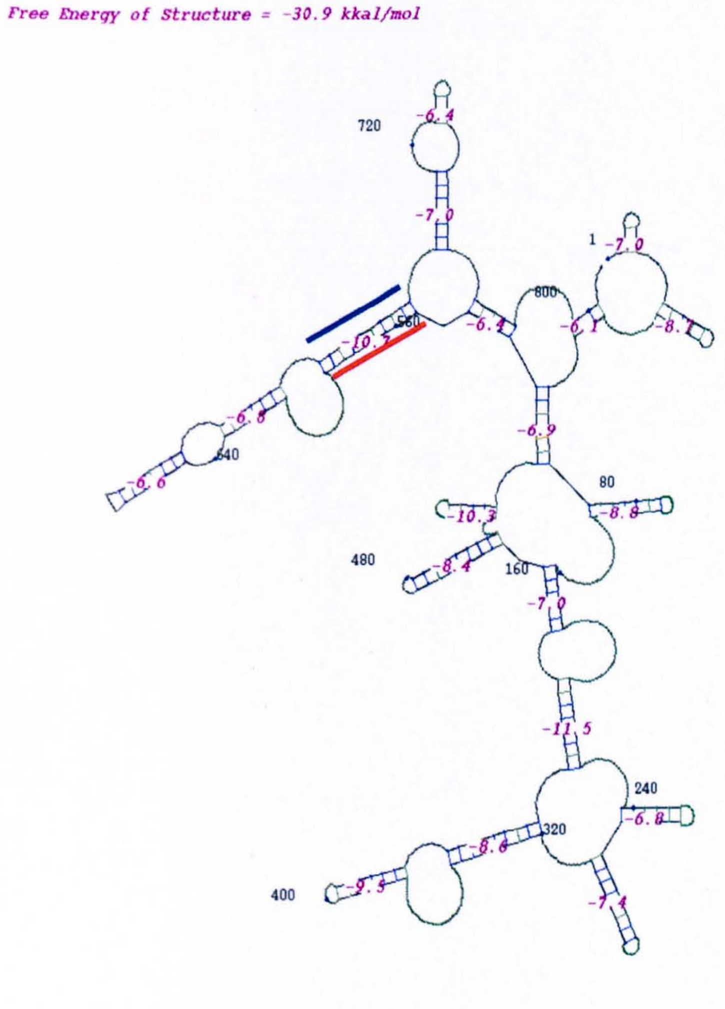
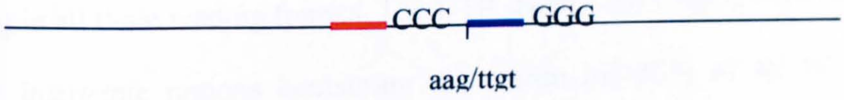


Figure 9.6 continued.....

c) A schematic of identified stems. The two complementary flanking regions are shown in letters (CCC and GGG) and the removed intron is indicated by the aag/ttgt-joining site resulting from splicing. In red and blue are indicated regions forming a stem. Below: Sequences of the red and blue regions are shown in order of appearances in bold. Splice site junctions are shown in underlined italics.

d) RNA structure for 110 5′ intergenic regions without the intron. On the figure is indicated which of the stems involved regions in the vicinity of the intron. As can be seen a stem was formed by regions located immediately before the CCC sequences and in the tgt acceptor-flanking sites. The free energy of the stem was –6.7 Kkal/mol, which was the third highest free energy in a stem loop in the entire structure.

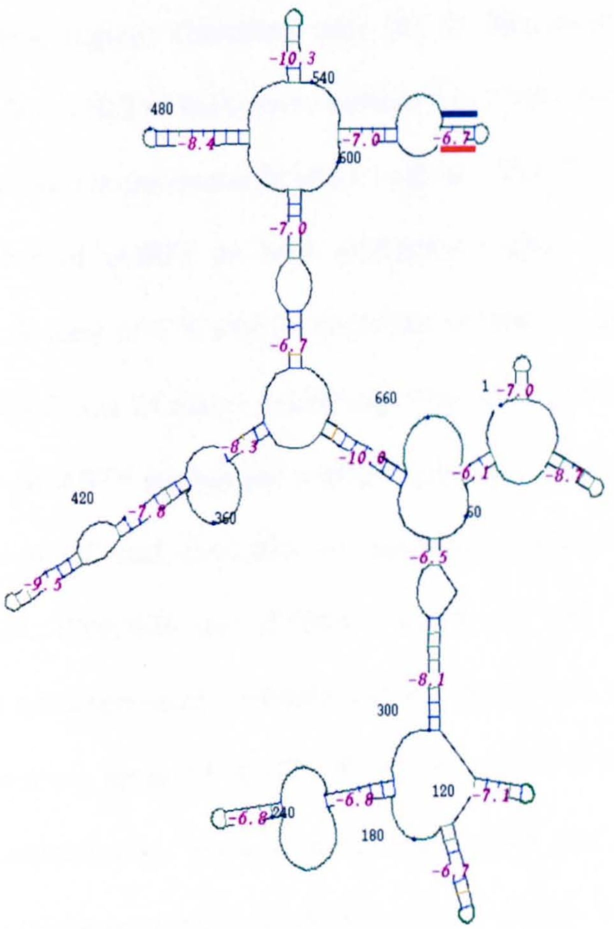
c)



ATTATTTTTATTTTTTAGATTTATATAGTACTTAAGTTTTAGAATTGAAA
TCACTAAAACAGAAGATATAAATATATTCCTTTTC**TATGT**TCAAAAATAC
GTTAAATTCATTATAAAAGTATATCCATTTTTATAATAAAGTTATCCAGAT
GCTAGTAAGAATAGCATTTTTGTATAAAATGCATCATATCTATATTTAAT
CAAAAATGTATTAAGCAACCCTCTATTT***AAGTTG***TTATATATAGGGTTAA
AGTATTTGCGTC**ACATA**TTGGGGACAACGTATATAAAGCAGCATGATTCT
ACTCAATTTACAAATTAAAAATAAAATCCCATTATA

d)

Free Energy of Structure = -32.2 kkal/mol

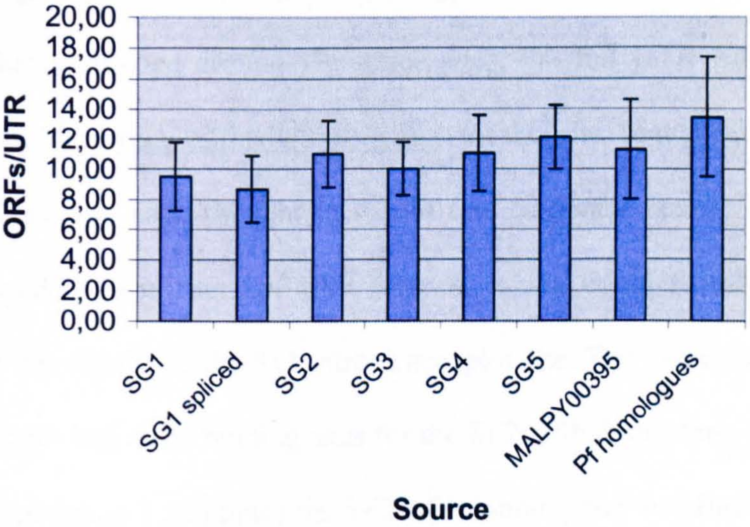


were analysed. Only the region from M2, which represented the longest 5'UTR observed (see Chapter VII), to immediately before the *yir* exon 1 was analysed. All uORFs starting with an ATG and containing at least 10 codons were found by scanning in all three reading frames. The UTR introns (Sv1 type) were removed from all SG1 intergenic regions containing this intron (67-80% of all SG1 intergenic regions, see above), and these regions were reanalysed for ORFs using the same criteria. On average between 9.5 (SG1) and 12.1 (SG5) uORFs were found per *yir* intergenic region. This was compared with 5'intergenic regions from *P.yoelii* homologues to *P.falciparum* genes with catalytic activity (13.4 uORFs per intergenic regions) and 10 genes located on the contig MALPY00395, found to be syntenic with *P.falciparum* chromosome 7 (Carlton, 2002). These had an average of 11.3 uORFs per intergenic region. Therefore, only the *yir* intergenic regions from SG1 (9.5 uORFs) and SG3 (10.3 uORFs) were outside what was observed from genes in syntenic and therefore chromosomal internal regions. For SG1, the UTR splicing reduced the number of uORFs on SG1 intergenic regions from 9.5 to 8.7 per intergenic region. A total of 470 uORFs contained at least 30 amino acids, and the largest identified ORF was 74 amino acids long. The ten largest uORFs (66 to 74 aa) were analysed by BLASTP against the annotated peptide database at www.tigr.org. Eight of the ten uORF had similarity to annotated proteins: three hypothetical proteins (PY06429, PY02636 and PY00442 at 48, 37 and 29 percent identity respectively), two occurrences of similarity to the same ribosomal protein (PY04916, 42% identity), three *yir* genes (PY07477, PY02140 and PY06937 at 64, 37 and 92 percent identity respectively). None of these were perfect matches, so this strongly suggest that there might be numerous ORFs in the *yir* intergenic regions originating from past recombination events with other genes.

Figure 9.7

Open reading frames in the intergenic regions and the effect of the UTR splicing

306 5′ intergenic regions (from Chapter VII) were analysed for the number of ORFs containing at least 10 amino acids. Only the regions from M2 (810 nt upstream of the ATG, see Chapter VII) and to the last nucleotides before the first *yir* exon were used. In addition 10 genes located on the longest *P.yoelii* contig, MALPY00395, which was highly syntenic to *P.falciparum* chromosome 7 (Carlton et. al., 2002), and 8 *P.yoelii* genes with homology to catalytic *P.falciparum* genes were also analysed for ORFs in their 5′ intergenic regions at the same distance from the first codon as the *yir* genes. The supergroup identity of these intergenic regions was established, and the average number of ORFs per intergenic region is shown. Sv1 type introns were removed from the SG1 sequences and the number of ORFs were calculated as described above. This was also plotted into the graph as the column SG1 spliced.


ORFs/UTR
20,00
18,00
16,00
14,00
12,00
10,00
8,00
6,00
4,00
2,00
0,00
SG1
SG1 spliced
SG2
SG3
SG4
SG5
MALPY00395
Pf homologues
Source

9.2.8 Exonic splicing enhancer predictions.

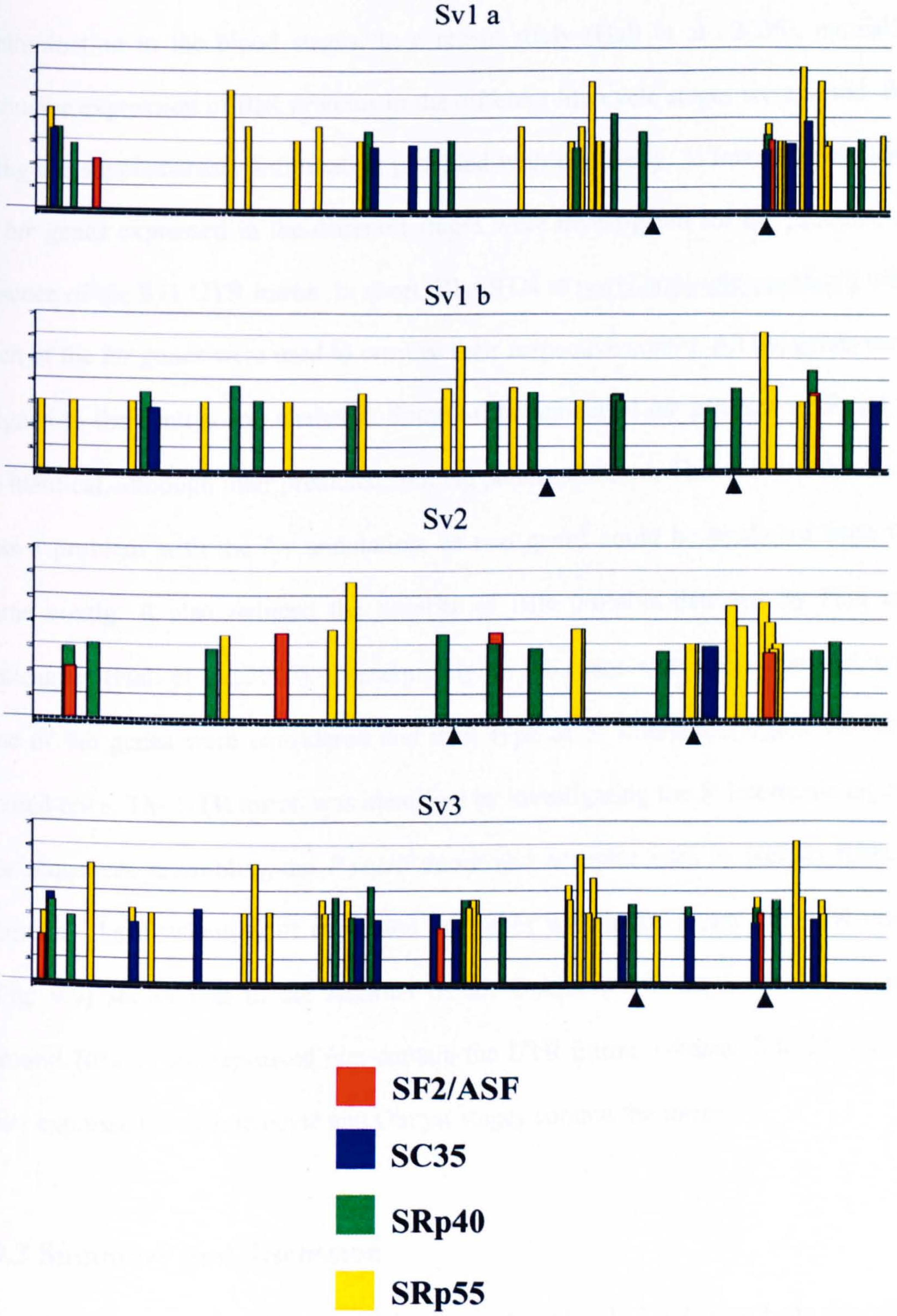
Although alternative splicing pathways are poorly understood at the moment, a general held idea is the exonic splicing enhancer regions (ESE) in the vicinity of intron donor and acceptor sites recruit various factors (Called SR proteins because of their high abundance of serine and arginine at their carboxy terminus, Brown et al., 2001), which could determine which splicing pathway is initiated (Brown et al., 2001).

To investigate if the Sv1, Sv2 and Sv3 types differed in the distribution of ESE regions they contained around the splice sites, the full UTR from each type was analysed at: http://rulai.cshl.edu/tools/ESE/. At this site, sequences were scanned for four ESE regions, each thought to recruit one particular factor. The analysis (Fig. 9.8) indicated that clusters of ESE regions, which could recruit all four factors, existed in the vicinity of the Sv1 introns acceptor site. The two splicing variants (Sv2 and Sv3) both had more binding sites for the SF2/ASF factor than the Sv1 types. For the Sv2 type (exon 1 skipping) the SF2/ASF binding site was further away from the intron acceptor site than for Sv1 a and Sv3, however Sv1 b had a similar SF2/ASF location with respect to the intron acceptor site. Overall, it is interesting to note that SF2/ASF binds only very few times in each UTR, but does so consistently in the vicinity of the introns acceptor site. It was investigated whether SF2/ASF homologues existed in *P.yoelii*. This was done by obtaining the SF2/ASF protein sequence from *Oryza sativa* (GI accession number 51854465) and performing a BLASTP against the annotated *P.yoelii* peptides at www.tigr.org. The highest scoring match was annotated as PY04347, "splicing factor, arginine/serine-rich", and had 42% identity to the input sequence and an e value of 1.4e-25. The identities suggested that this could be the *P.yoelii* homologue of SF2/ASF.

Figure 9.8

Exonic splicing enhancer predictions

Experimentally verified UTRs from two Sv1 (the two most divergent Sv1 types: Sv1 a and Sv1 b) and the Sv2 and Sv3 types was uploaded to the exonic enhancer prediction programme at: http://rulai.cshl.edu/tools/ESE/. This program scans sequences for binding sites for the four exonic enhancer proteins: S2F/ASF, SC35, SRp40 and SRp55. The relative strength of these predictions is indicated by bar height in the plots. On each plot, the location of the (Sv1 type) intron is indicated. For all sequences, predictions of exonic enhancers were clustered around the introns donor and acceptor sites. Also, in all predictions, S2F/ASF was the least frequent, however, it was predicted to bind in the vicinity of the acceptor site for the Sv1 intron.


Sv1 a
Sv1 b
Sv2
Sv3
SF2/ASF
SC35
SRp40
SRp55

9.2.9 UTR introns in stage expressed BIR proteins

The low abundance of Sv2 and Sv3 splicing types could reflect gametocyte contamination in the blood stages. In a recent study (Hall et al., 2005), mutually exclusive expression of BIR proteins in the different life cycle stages were found. By using the supplementary information provided with this study, 5′ intergenic regions of *bir* genes expressed in the different stages were investigated for the presence or absence of the Sv1 UTR intron. In short, BLASTN at www.http://plasmodb.org with each of the *bir* genes were used to retrieve their respective contig. All the genes were aligned to the contigs and analysed. Some of the annotated *bir* genes were found to be identical, although their predicted splicing pattern differed. This showed that there was a problem with the *bir* annotation, as two genes could be predicted from the same contig. It also reduced the number of BIR proteins detected by Hall and colleagues (Hall et al., 2005). Consequently, in the cases where this occurred, only one of the genes were considered and their type of 5′ intergenic region was only scored once. The UTR intron was identified by investigating the 5′ intergenic regions for sequences resembling the *P.yoelii* donor and acceptor sites as well as flanking regions. The distribution of expressed *bir* genes with and without the UTR intron (Fig. 9.9) shows that in the Asexual blood, Ookinete and the Sporozoite stages, around 70% of the expressed *birs* contain the UTR intron, whereas 0 to 22% of the *birs* expressed the gametocyte and Oocyst stages contain the intron.

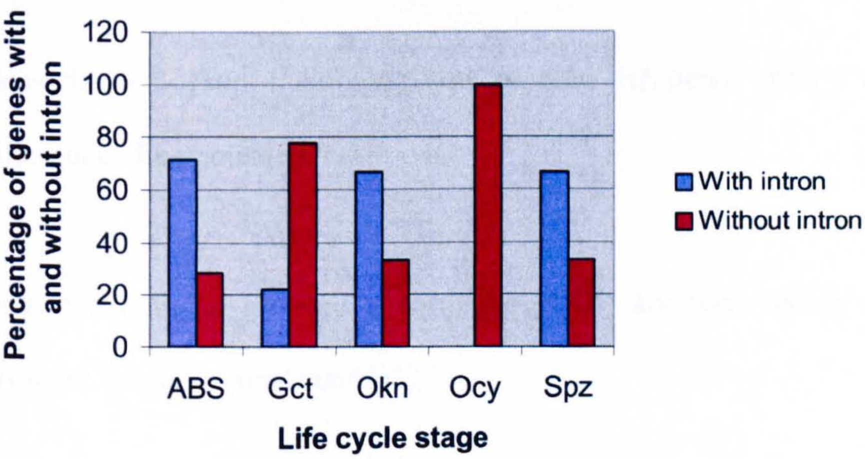
9.3 Summary and discussion

In this Chapter, three different splicing variants of the 5’ UTR have been described. One type of splicing (Sv1) was predominant and occurred between two complementary flanking regions around 110 nt upstream of the *yir* exon 1.

Figure 9.9

Distribution of UTR intron in stage expressed *bir* genes

The *bir* genes expressed in the different life cycle stages were analysed for the presence or absence of the UTR intron in their 5′ intergenic regions. A total of 29 *bir* genes could be analysed, and the number of these containing the intron or not in each stage were: Asexual blood stages (ABS): 5/2, Gametocyte (Gct): 2/7, Ookinete (Okn): 4/2, Oocyst (Ocy): 0/4, Sporozoite (Spz): 2/1. This is presented as the percentages of genes with or without the intron.


Percentage of genes with and without intron
120
100
80
60
40
20
0
ABS
Gct
Okn
Ocy
Spz
Life cycle stage
🟦 With intron
🟥 Without intron

Two minor splice variants (Sv2 and Sv3) led to a *yir* exon 1 skipping event and removal of a large upstream intron in the UTR respectively. Sv2 used the same intronic donor site as Sv1, whereas Sv3 used the same acceptor site as Sv1. Some *yir* UTRs did not contain the intron at all, however, the Sv1 intron was distinctly distributed among the supergroups with 67-80% of SG1, and 23% of SG2- and 4.8% of SG5-genes containing the intron in their UTRs.

Assuming that this splicing has a function, four different hypothesises for what this could be were considered further (in increasing order of likelihood):

a) Does the Sv2 exon 1 skipping type produce *yir* genes with a distinctly different coding potential?

b) Does the splicing produce a particular RNA structure, which could be involved in translational control?

c) Does the splicing increase the likelihood for translation by removing uORFs?

d) Is the splicing a way to ensure stage specific expression of YIR proteins?

a) The minor Sv2 splicing type affected the coding potential of the *yir* mRNA by skipping the first exon, and in two annotated *yir* genes, this splicing type had been identified by the automated prediction which joined an alternative first exon to the second *yir* exon. An additional 19 *yir* genes were found to contain highly similar and translatable in-frame alternative first exons. It thus remains theoretically possible that this splicing can generate *yir* mRNAs with an alternative coding potential. However, the fact that the intron would lead to truncated *yir* transcripts in the vast majority of the cases (because 819 *yir* genes did not contain an alternative in-frame first exon) makes it unlikely that this mechanism has evolved for the purpose of only a tiny fraction of the genes spliced by this.

b) The removal of the intron lead to the formation of a structure resembling an IRES. IRES are involved in cap-independent translation (Seino et al., 2005); however, they could also function by bypassing uORFs (Almeida et al., 2005). However, the existence of cellular IRES is disputed (Kozak et al., 2005), and there are several *caveats* to RNA folding predictions at the moment. Most importantly, although the most thermodynamically favourable structure can be predicted computationally, *in vivo* factors such as RNA binding proteins, chemical modifications and other processes are ignored by all current RNA prediction algorithms (Gardner et al., 2004). In this case, it can just be concluded that interactions between the two complementary regions are not thermodynamically favourable, and that splicing leads to a structure resembling an IRES. However, because of the above-mentioned *caveats*, it is not possible to say if the RNA molecule assumes this structure *in vivo*, and much less to say if the proposed IRES would have any function.

c) The presence of uORFs are thought to decrease the likelihood of translation due to premature dissociation of the ribosome (as reviewed by Mignone et al., 2002 and Pesole et al., 2001). For SG1 9.5 upstream open reading frames (uORFs) and for SG5 12.1 uORFs were identified per gene from the different supergroups. This rather large number of uORFs could decrease the efficiency, with which a *yir* mRNA is translated. The Sv1 type splicing only reduced the average number of uORFs by 0.8 per intergenic region in SG1. This still leaves the translational apparatus to deal with 8.7 uORFs per intergenic region, so it is thought to be less likely that the splicing has evolved to function this way. However, the possible IRES (see b) could play a role in bypassing these uORFs, as this would allow the translational apparatus to assemble at a short distance away from the ATG start codon. The analysis clearly showed that these uORFs originated from past recombination events, where random parts of genes had been inserted into the *yir* intergenic sequences. It appears less likely that these uORFs could come to resemble annotated genes by random mutations alone. This would pose an increasing problem for the translational apparatus. However, it has been suggested that the translational apparatus can bypass most uORF because they lack a Kozak consensus sequence (as reviewed by Mignone et al., 2002 and Pesole et al., 2001). Since it is not known if *P.yoelii* uses a Kozak sequence under normal circumstances, it cannot be estimated how much of a problem these uORFs play in *P.yoelii.*

d) The location of one particular region resembling an exonic splicing enhancer region at the 3′ end of the Sv1 (and Sv3 intron) was interesting. This region could recruit the *P.yoelii* homologue of SF2/ASF. In terms of alternative splicing, SF2/ASF has been found to promote the usage of the most proximal 3′ splice site, but a more distal 3′ splice site is used when another factor (hnRNP A1) is present and antagonizes the effect of SF2/ASF (Bai et al., 1999). It has also been found; that tissue specific expression of these two factors can govern which alternative 3′ splice site is used (Pollard et al., 2001). As only four factors were investigated through this analysis, it remains possible that other, unknown factors could also play an important role. In this study, the low abundance of Sv2 and Sv3 types could originate from gametocytes present in the blood stages. Indeed, differential splicing in the gametocyte stages have been observed for the *b7* and *stevor* transcripts (Pace et al., 1998 and Sutherland et al., 2001), and it could be that different splicing/ESE binding factors, are present in the gametocyte. The distribution of expressed *bir* genes (BIR proteins) with and without the UTR intron (Fig. 9.9) shows that in the Asexual blood (ABS), Ookinete and the Sporozoite stages, around 70% of the expressed *birs* contain the UTR intron, whereas only 0 to 22% of the *birs* expressed in the Gametocyte (Gct) and Oocyst stages contain the intron. It was only possible to investigate a relatively low number of expressed *bir* genes (29 in total), so this limits the representatively of this analysis. However, in ABS and Gct, the proportion of expressed genes containing the intron was reversed. If the gametocyte promotes the Sv2 type splicing over the Sv1 type through the use of different (or antagonizing) splicing factors, this could prevent translation of these truncated mRNAs.

This explanation is favoured, as biologically; such a mechanism would be very sensible. Since the gametocytes are in circulation along with the asexual blood stage parasites, host antibody responses against the set of YIR proteins expressed in the asexual blood stages, would also target the gametocyte.

If, through alternative splicing, mediated by different expression of splicing components, exon 1 skipping occurs more frequently in the gametocytes, YIR proteins encoded by SG1 might not be translated. This would ensure that the gametocyte expresses a different set of YIR proteins than the asexual blood stages. Since transmission can be seen as the largest bottleneck of a successful infection, there would be strong evolutionary forces operating to optimise the likelihood for a successful transmission. This is just a guess at a possible function. In order to investigate if this guess is correct, several experiments has to be undertaken, which is described in more detail in Chapter X.

Chapter X

Conclusions and future perspectives

10.1 Conclusions

The genes within the *yir* multigene family in *P.yoelii* are organized into five supergroups, SG1 to SG5 (Fig. 10.1 a). Four of these supergroups (SG2 to SG5) are distinctively different from each other, while the largest supergroup (SG1) is composed of more heterogeneous genes. Each of the supergroups has a distinct distribution on annotated subtelomeric contigs (Fig. 10.1 b), suggesting they are located in different chromosomal regions. SG1 is co-localized with members of all the other four supergroups, as it existed on shared contigs with genes from all other supergroups, which suggest that this supergroup is located throughout the chromosomal regions containing *yir*. As a model for how the supergroups are thought to be located (Fig. 10.1 c), SG1 is present with all the remaining supergroups all over the *yir*-containing chromosome ends.

The 5′ intergenic regions (Fig. 10.1 d) is divided into distinct sets, which are each localized in front of *yir* genes from a distinct supergroup, whereas the two sets of 3′ intergenic regions, identified in this study, do not follow the supergroups.

These findings could indicate that the maintenance of distinct supergroups is probably kept by separating them into distinct chromosomal regions, where intra-supergroup recombinations are favoured over inter-supergroup recombinations. It can also be speculated why the 5′intergenic regions reflects the supergroups and the 3′ intergenic regions do not. This could either suggest two transcendent levels of regulation: one transcriptional (5′UTR) and one post-transcriptional (3′UTR), or that only one of the intergenic regions is important for regulation of expression. However, the presence of UTR introns in some of the 5′intergenic regions suggest that the 5′ UTR could also be involved in post transcriptional regulation.

Figure 10.1

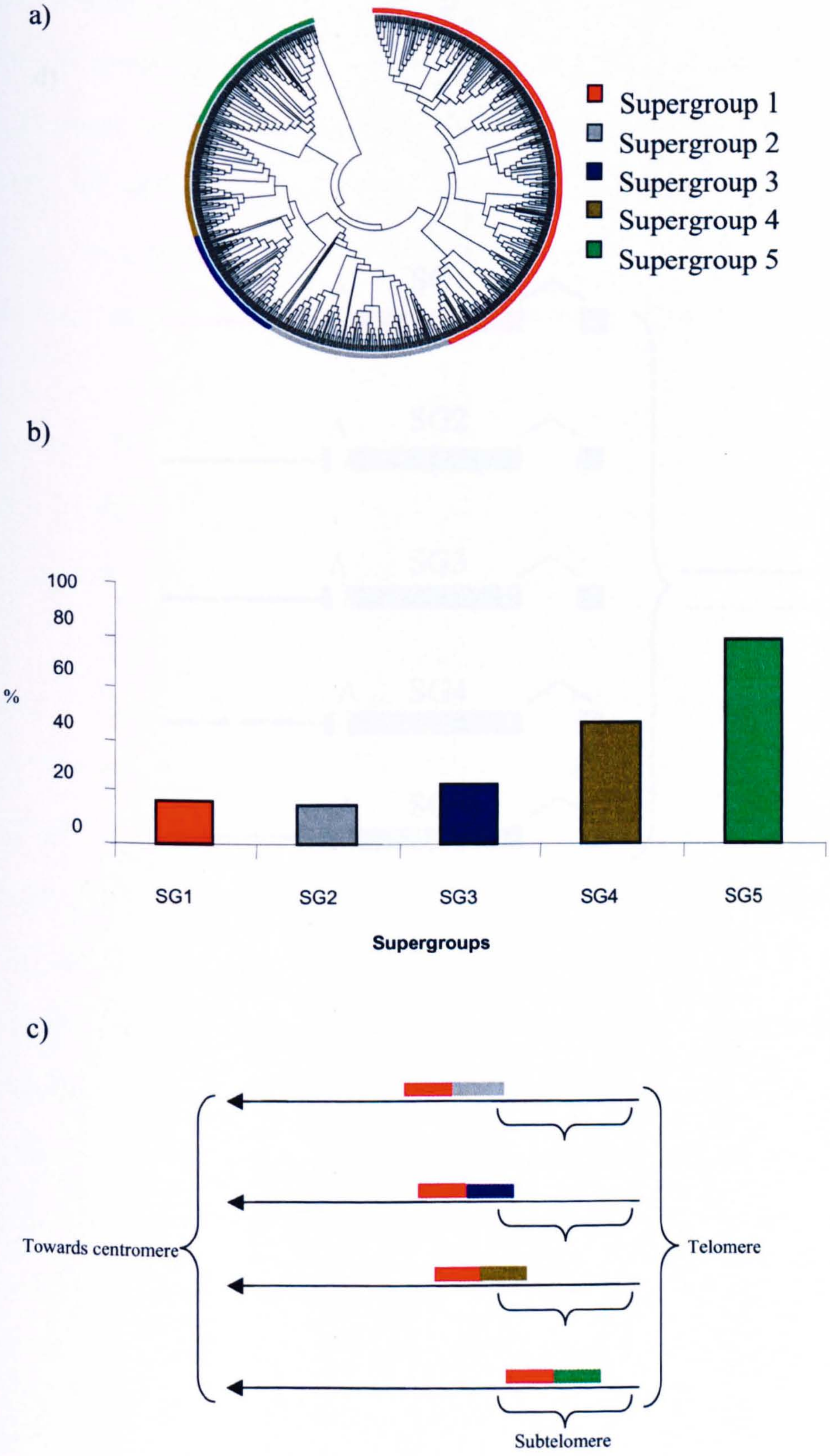
Organization of *yir* genes and intergenic regions

a) Phylogenetic NJ tree of the five *yir* genes supergroups

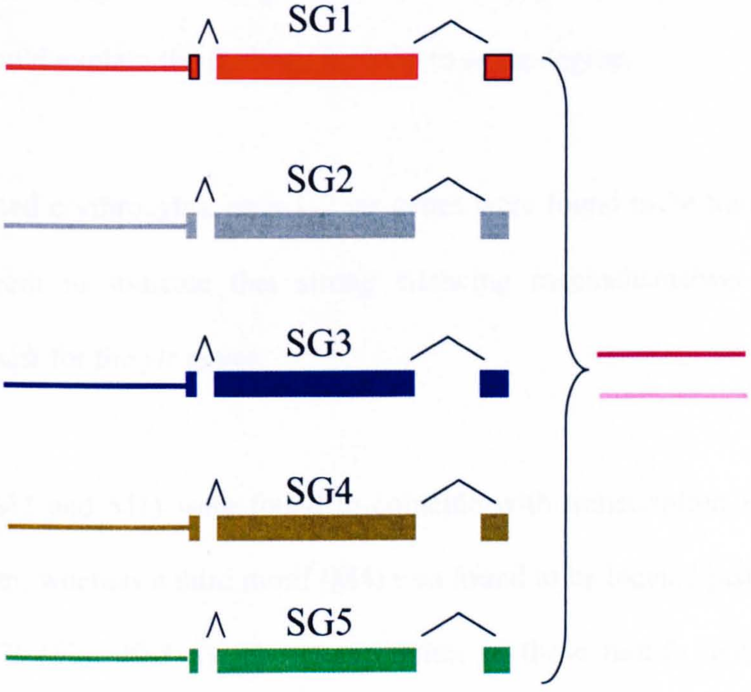
b) The percentages of each of the five supergroups located on the annotated subtelomeric contigs.

c) Proposed chromosomal organization of the *yir* genes on five hypothetical chromosomes. SG1 (red) was co-localized on the same contigs with genes from all other supergroups. Placing SG1 genes adjacent to all other supergroups indicates this. A high proportion of SG1 genes co-localized with SG5 genes on the subtelomeric contigs, which is also indicated.

d) Co-occurrences of intergenic sets and *yir* supergroups. The 5′intergenic regions existed in discrete sets that were located in front of *yir* genes from a particular supergroup. The 3′ intergenic regions existed in two discrete sets, but these were not localized after *yir* genes from the supergroups in any systematic manner.


a)
🟥 Supergroup 1
▨ Supergroup 2
■ Supergroup 3
■ Supergroup 4
🟩 Supergroup 5
b)
100
80
60
40
20
0
%
SG1
SG2
SG3
SG4
SG5
Supergroups
c)
Towards centromere
Telomere
Subtelomere

d)



Supergroups located primarily at chromosome ends could be regulated differentially by epigenetic mechanisms. Since 5′ intergenic regions are important for transcriptional regulation, this would therefore be a way to differentially control the transcription of the different *yir* supergroups. More transcripts were detected from SG1, SG2 and SG3 (Fig. 10.2) than the highly subtelomerically located SG4 and SG5. This could imply that more genes in SG4 are epigenetically silenced, whereas cloning bias could explain the findings for SG5 to some degree.

In single infected erythrocytes, only 1-2 *yir* genes were found to be transcribed, and this would seem to indicate that strong silencing mechanisms/weak activation mechanisms exist for the *yir* genes.

Two motifs (M2 and M1) were found to coincide with transcription initiation and polyadenylation, whereas a third motif (M4) was found to be located just upstream of the *yir* 5′ UTR (Fig. 10.3 a). The conservation of these motifs in (Fig. 10.3 b) showed that while M4 was completely conserved in front of all analysed *yir* genes, M2 and M1 were less conserved. Although there was some correlation between distribution in the supergroups of M2 and M1 and detected transcripts, this was not absolutely clear-cut. It is therefore thought that M2 and M1 can be used frequently as transcription initiation and polyadenylation sites respectively. However, since only polymerase III uses internal transcription initiation sites (Brown, 2002) is less clear what the role of M2 could be. In contrast to this, the universality of M4 would suggest a universal role in transcriptional regulation of all *yir* genes, regardless of supergroup. Transient transfections did not elucidate the function of M4, but it is possible that *yir* transcription is at such a low level that other approaches are needed to determine this. However, a low level of reporter gene transcription was observed.

Figure 10.2

Transcription profile of *yir*

The numbers of transcripts detected from all the supergroups are indicated. These were all detected in the same blood sample, grown in an immunocompromised mouse, and therefore they reflect a sample of the *yir* repertoire in the absence of any immune mechanisms.

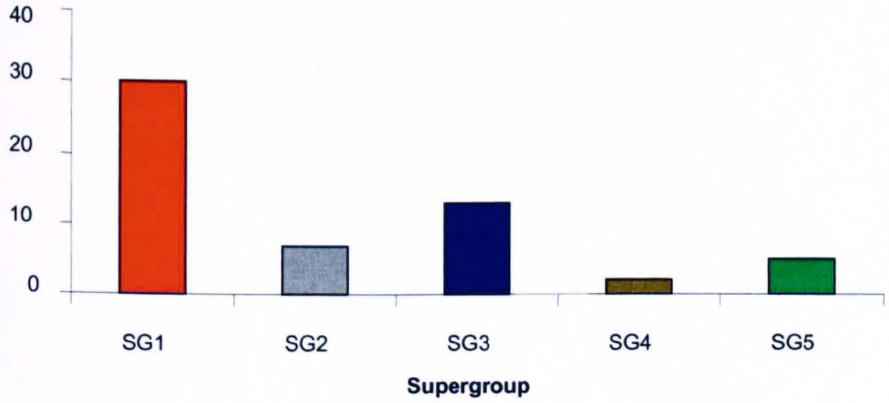
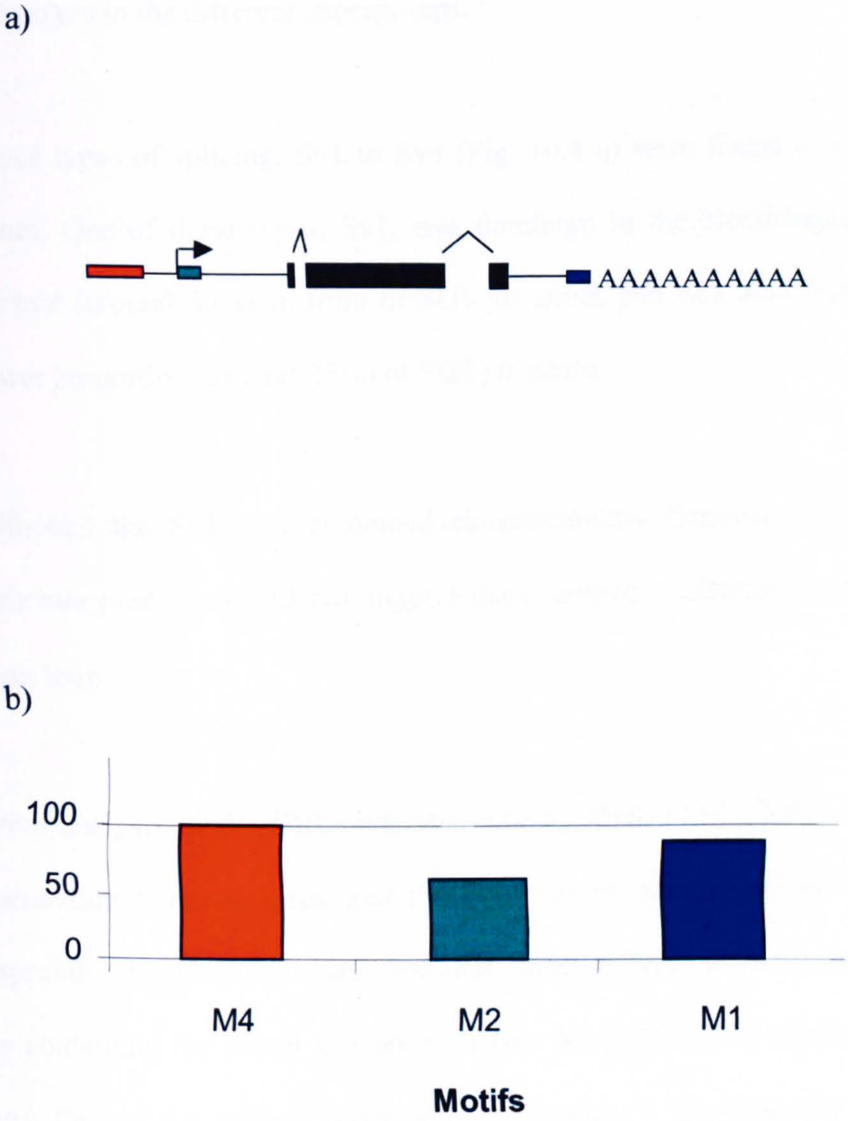

40
30
20
10
0
SG1
SG2
SG3
SG4
SG5
Supergroup

Figure 10.3

Putative transcription regulatory motifs

a) The transcription start site for (some) *yir* genes were located some 772 nt upstream of the ATG, in a semi-conserved motif (green, M2). An additional 138 nt upstream from this, another highly conserved motif (red, M4) was located. No transcripts were found to contain this motif, showing that it is located outside the transcribed region. A third highly conserved motif (blue, M1) coincided with polyadenylation.

b) Conservation of motifs among *yir* genes. The three motifs are indicated by their colour coding (red (M4), green (M2) and blue (M1)). The percentages of analysed intergenic regions containing each of these motifs were calculated and are shown.


a)
AAAAAAAAAA
b)
100
50
0
M4
M2
M1
Motifs

Therefore, in terms of transcriptional regulation, it is thought that all *yir* genes interact with the same set of regulatory factors indiscriminately. However, epigenetic mechanisms are very likely to affect how efficiently these factors can access *yir* promoters in the different supergroups.

Three types of splicing, Sv1 to Sv3 (Fig. 10.4 a) were found to occur for SG1 *yir* genes. One of these types, Sv1, was dominant in the bloodstages and was highly present (around 80%) in front of SG1 *yir* genes and was also present in front of a lower proportion (around 25%) of SG2 *yir* genes.

Although the Sv1 type contained complementary flanking regions, global RNA structure predictions did not suggest these formed a thermodynamically favourable stem loop.

From analysis of the BIR proteome analysis (Hall et al., 2005), there was a clear correlation between stage and the presence or absence of the Sv1 intron type; especially the gametocyte and Ookinete stages differed from all the remaining stages by containing the intron in front of a very low (0 to 22%) proportion of expressed *birs*. One of the splicing types, Sv2, led to exon 1 skipping, and if Sv2 splicing is prevalent in the gametocyte stage, it is possible that these genes would not be translated at all.

Predictions of putative exonic splicing enhancer regions surrounding the Sv1 type intron led to the identification of a binding site for the SF2/ASF splicing enhancer close to the Sv1 introns acceptor site (Fig. 10.4 b).

Figure 10.4

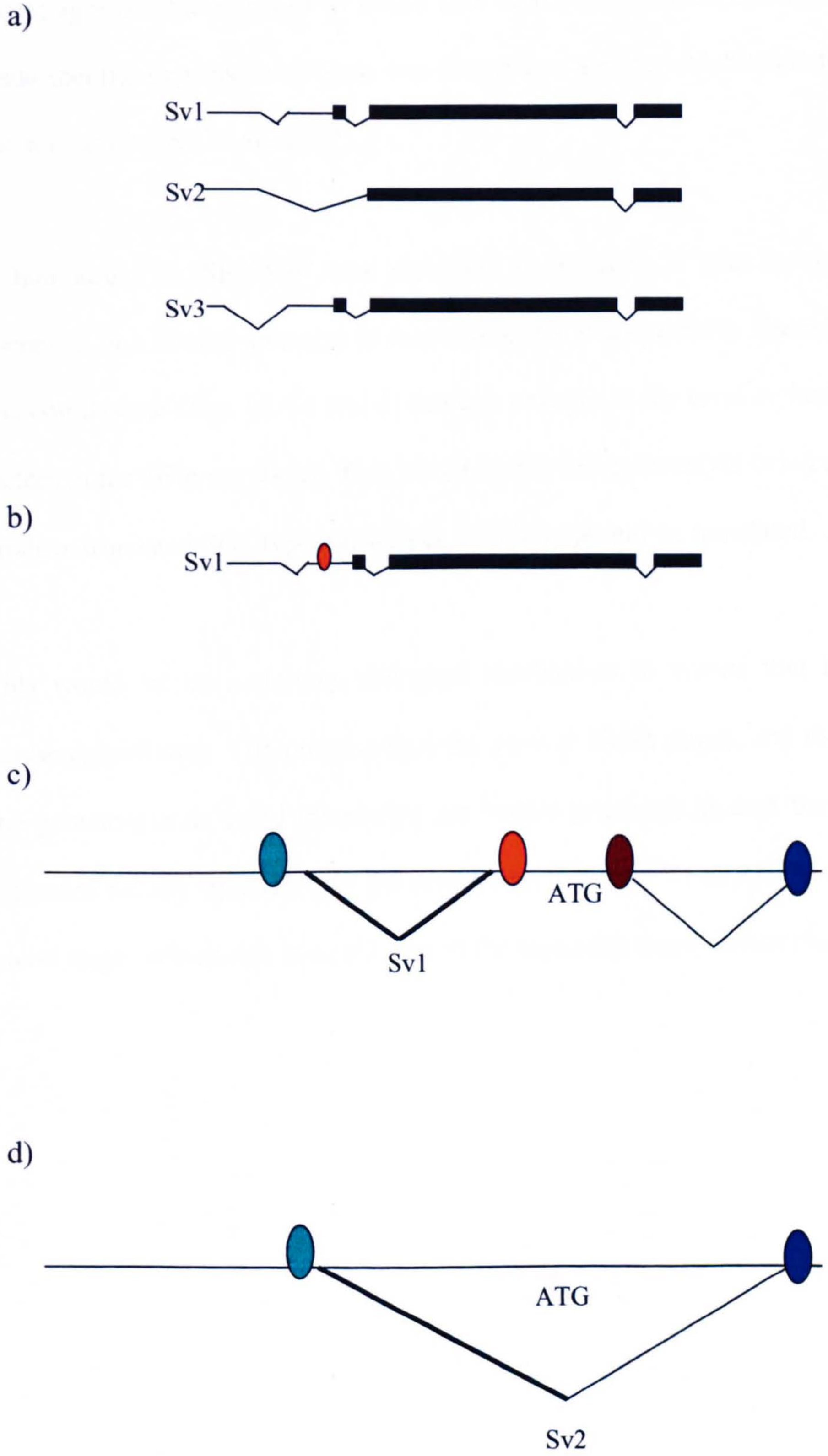
Proposed mechanism of alternative splicing

a) Three alternatively spliced types of *yir* genes (Sv1, Sv2 andSv3) were identified. For two of these (Sv1 and Sv2), this splicing occurred only in the *yir* 5′UTR, whereas another (Sv2) led to skipping of the first *yir* exon. Sv1 was much more prevalent in the blood stages than Sv2 and Sv3.

b) In the Sv1 type of splicing, a rarely occurring splicing enhancer factor was predicted to bind consistently downstream of the intron acceptor site. This factor, SF2/ASF has been shown to be involved in alternative 3′splicing selection.

c) In some stages (e.g. blood stages), the Sv1 type of splicing occurs as four hypothetical splicing enhancer factors delimits the Sv1 intron and the *yir* intron 1 and splicing occurs separately.

d) In other stages (e.g. gametocyte) stage, the Sv1 intron 3′ splicing enhancer and/or the 5′ intron 1 enhancer is prevented from binding. This results in the Sv1 type intron donor site being joined with the *yir* intron 1 acceptor site, leading to the Sv2 type of splicing, skipping exon 1.


a)
Sv1
Sv2
Sv3
b)
Sv1
c)
ATG
Sv1
d)
ATG
Sv2

SF2/ASF has been found to promote the usage of the most proximal 3′ splice site, but a more distal 3′ splice site is used when another factor (hnRNP A1) is present and antagonizes the effect of SF2/ASF (Bai et al., 1999). It has also been found; that tissue specific expression of these two factors can govern which alternative 3′ splice site is used (Pollard et al., 2001).

A homologue to SF2/ASF was identified in *P.yoelii*. If this factor (or another operating in a similar manner) is responsible for stage specific alternative splicing, this could occur (Fig. 10.4 c and d) through changes in the level or function of these factors in the different stages. This would lead to the gametocyte being more likely to produce truncated SG1 type transcripts, which might not be translated.

This would be an important biological mechanism to ensure that the gametoyte expressed different YIR proteins than the asexual blood stages, and therefore allows the gametocyte to avoid circulating antibodies generated against the YIR proteins (encoded mainly by SG1 type transcripts, see Chapter V) expressed in the asexual blood stage. Which role it would play in the mosquito stages is less clear however.

To summarize, the conclusions from this work:

1. *Yir* genes belong to five supergroups with distinct gene sizes and subtelomeric localisations.
2. In the absence of immune-mechanisms, transcription occurs from all supergroups, and SG1 and SG3 were more transcriptionally active than SG4.
3. In single infected RBCs at the Schizont stage, only 1-2 *yir* genes are transcribed.
4. *Yir* transcription initiates 772 nt upstream of the ATG and terminates 630 nt downstream of the translational stop codon.
5. *Yir* genes are polyadenylated, and this frequently occurred at an unusual and conserved triple-repeat (*CGA*x3*)* motif (M1).
6. Transcription initiated at a highly conserved motif (M2), some 140 nt downstream of a universally conserved triple-repeat (*TCTCTC*) motif (M4).
7. M4 was located outside the transcribed region of *yir* genes.
8. One of the *yir* supergroups, SG1, contained three introns in the UTR, one of which led to skipping of exon 1.
9. A rarely occurring exonic splicing enhancer, with a proven role in 3′ alternative splicing, was predicted to bind to regions in the UTR introns 3′ splice site. A homologue to this factor (SF2/ASF) was identified in *P.yoelii.*
10. A distinct difference in *P.berghei* BIR expression was found with respect to the UTR introns; asexual blood stages expressed several BIR proteins with this intron, whereas only a few were expressed in the gametocytes stage.

10.2 Future perspectives

I will here try to summarize which experiments would be important in my opinion to take this project further.

First of all, in this study, *yir* transcription was measured by RT-PCR. Although this gave some indications of the extent and diversity of the transcribed repertoire, this is not quantifiable by any means. Microarrays containing *yir* probes were being developed at the end of this project, but unfortunately too late to take advantage of this. With these, it will be very important to find out how much transcription occurs from the different supergroups, not only in the asexual blood stages, but also in particular in the gametocyte stage. This would give an indication of how much epigenetic regulation occurs in the different supergroups. Another important question that could be assessed through microarray analysis is how quickly a large number of *yir* genes are transcribed during an infection. For this, single cell infections, followed by microarray analysis at different time points could be used to measure this dynamic.

In my opinion, it would also be important to elucidate the function of M4. As discussed (in Chapter VIII) there could be some technical reasons for the outcome of the present transfections studies, but it could also be that the likelihood for transcription would have been low in any case. Taking the technical considerations into account, it would maybe be better if the *yir* intergenic region was cloned in front of a drug resistance gene while also maintaining a different drug resistance gene in the construct. This would allow for selection of the possible low proportion of parasites where the promoter is active, and thus allow quantification of the effects of

M4. In addition, it should also be attempted to identify nuclear proteins interacting with M4 through gel shift experiments.

Another question that would have to be assessed in a quantifiable manner, where microarray could also be useful, was the levels of Sv1, Sv2 and Sv2 types of splicing in the different stages. Since relatively little is known about the roles of different spliceosomal component in general, and even less in *Plasmodium*, the Sv1 to Sv3 types of splicing could be a starting point (if stage specific splicing does occur) for trying to identify differentially regulated splicing factors in *Plasmodium yoelii.* If this is indeed the case, it would be important to investigate which splicing factors are differentially expressed in a manner, which could qualify them as candidates for alternative splicing. This thesis suggests that one possible candidate would be the *P.yoelii* homologue of SF2/ASF2.

In this respect, over expression of this candidate by stable or transient transfections, could be used to assess the effect on the proportion of Sv1-Sv3 types of splicing by the above-mentioned microarray. It could also be speculated if knocking out of this splicing factor would compromise the proposed secret life of gametocytes and hence transmission?

References

Abdel-Latif, M. S., K. Dietz, et al. (2003). "Antibodies to Plasmodium falciparum rifin proteins are associated with rapid parasite clearance and asymptomatic infections." Infect Immun **71**(11): 6229-33.

Bai, Y., D. Lee, et al. (1999). "Control of 3' splice site choice in vivo by ASF/SF2 and hnRNP A1." Nucleic Acids Res **27**(4): 1126-34.

Barnes, C. A., M. M. MacKenzie, et al. (1995). "Efficient translation of an SSA1-derived heat-shock mRNA in yeast cells limited for cap-binding protein and eIF-4F." Mol Gen Genet **246**(5): 619-27.

Balaji, S., M. M. Babu, et al. (2005). "Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains." Nucleic Acids Res **33**(13): 3994-4006.

Baruch, D. I., B. L. Pasloske, et al. (1995). "Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes." Cell **82**(1): 77-87.

Bell, A. C., A. G. West, et al. (2001). "Insulators and boundaries: versatile regulatory elements in the eukaryotic." Science **291**(5503): 447-50.

Bentley, D. (2002). "The mRNA assembly line: transcription and processing machines in the same factory." Curr Opin Cell Biol **14**(3): 336-42.

Biggs, B. A., L. Gooze, et al. (1991). "Antigenic variation in Plasmodium falciparum." Proc Natl Acad Sci U S A **88**(20): 9171-4.

Blackman, M. J., H. G. Heidrich, et al. (1990). "A single fragment of a malaria merozoite surface protein remains on the parasite during red cell invasion and is the target of invasion-inhibiting antibodies." J Exp Med **172**(1): 379-82.

Blythe, J. E., T. Surentheran, et al. (2004). "STEVOR--a multifunctional protein?" Mol Biochem Parasitol **134**(1): 11-5.

Boschet, C., M. Gissot, et al. (2004). "Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 Plasmodium falciparum clones." Mol Biochem Parasitol **138**(1): 159-63.

Bottius, E., N. Bakhsis, et al. (1998). "Plasmodium falciparum telomerase: de novo telomere addition to telomeric and nontelomeric sequences and role in chromosome healing." Mol Cell Biol **18**(2): 919-25.

Bozdech, Z., M. Llinas, et al. (2003). "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum." PLoS Biol **1**(1): E5.

Breman, J. G. (2001). "The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden." Am J Trop Med Hyg **64**(1-2 Suppl): 1-11.

Breman, J. G., M. S. Alilio, et al. (2004). "Conquering the intolerable burden of malaria: what's new, what's needed: a summary." Am J Trop Med Hyg **71**(2 Suppl): 1-15.

Bull, P. C., B. S. Lowe, et al. (1999). "Antibody recognition of Plasmodium falciparum erythrocyte surface antigens in Kenya: evidence for rare and prevalent variants." Infect Immun **67**(2): 733-9.

Calderwood, M. S., L. Gannoun-Zaki, et al. (2003). "Plasmodium falciparum var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron." J Biol Chem **278**(36): 34125-32.

Cann, H., S. V. Brown, et al. (2004). "3' UTR signals necessary for expression of the Plasmodium gallinaceum ookinete protein, Pgs28, share similarities with those of yeast and plants." Mol Biochem Parasitol **137**(2): 239-45.

Carlton, J., J. Silva, et al. (2005). "The genome of model malaria parasites, and comparative genomics." Curr Issues Mol Biol **7**(1): 23-37.

Carlton, J. M., S. V. Angiuoli, et al. (2002). "Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii." Nature **419**(6906): 512-9.

Chen, L. and J. Widom (2005). "Mechanism of transcriptional silencing in yeast." Cell **120**(1): 37-48.

Chen, Q., V. Fernandez, et al. (1998). "Developmental selection of var gene expression in Plasmodium falciparum." Nature **394**(6691): 392-5.

Chen, Q., A. Heddini, et al. (2000). "The semiconserved head structure of Plasmodium falciparum erythrocyte membrane protein 1 mediates binding to multiple independent host receptors." J Exp Med **192**(1): 1-10.

Chen, Q., F. Pettersson, et al. (2004). "Immunization with PfEMP1-DBL1alpha generates antibodies that disrupt rosettes and protect against the sequestration of Plasmodium falciparum-infected erythrocytes." Vaccine **22**(21-22): 2701-12.

Cheng, Q., N. Cloonan, et al. (1998). "stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens." Mol Biochem Parasitol **97**(1-2): 161-76.

Corredor, V., E. V. Meyer, et al. (2004). "A SICAvar switching event in Plasmodium knowlesi is associated with the DNA rearrangement of conserved 3' non-coding sequences." Mol Biochem Parasitol **138**(1): 37-49.

Coulson, R. M., N. Hall, et al. (2004). "Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum." Genome Res **14**(8): 1548-54.

Cunningham, D. A., W. Jarra, et al. (2005). "Host immunity modulates transcriptional changes in a multigene family (yir) of rodent malaria." Mol Microbiol **58**(3): 636-47.

Daily, J. P., K. G. Le Roch, et al. (2005). "In vivo transcriptome of Plasmodium falciparum reveals overexpression of transcripts that encode surface proteins." J Infect Dis **191**(7): 1196-203.

De Las Penas, A., S. J. Pan, et al. (2003). "Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing." Genes Dev **17**(18): 2245-58.

Deitsch, K. W., M. S. Calderwood, et al. (2001). "Malaria. Cooperative silencing elements in var genes." Nature **412**(6850): 875-6.

Deitsch, K. W., A. del Pinal, et al. (1999). "Intra-cluster recombination and var transcription switches in the antigenic variation of Plasmodium falciparum." Mol Biochem Parasitol **101**(1-2): 107-16.

del Portillo, H. A., C. Fernandez-Becerra, et al. (2001). "A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax." Nature **410**(6830): 839-42.

del Portillo, H. A., M. Lanzer, et al. (2004). "Variant genes and the spleen in Plasmodium vivax malaria." Int J Parasitol **34**(13-14): 1547-54.

Donze, D. and R. T. Kamakaka (2001). "RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae." Embo J **20**(3): 520-31.

Duffy, M. F., T. J. Byrne, et al. (2005). "Broad analysis reveals a consistent pattern of var gene transcription in Plasmodium falciparum repeatedly selected for a defined adhesion phenotype." Mol Microbiol **56**(3): 774-88.

Duraisingh, M. T., T. S. Voss, et al. (2005). "Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum." Cell **121**(1): 13-24.

English, M., C.R. Newton (2002). “Malaria: pathogenicity and disease.” Chem Immunol **80**: 50-56

Engwerda, C. R., L. Beattie, et al. (2005). "The importance of the spleen in malaria." Trends Parasitol **21**(2): 75-80.

Escalante, A. A., O. E. Cornejo, et al. (2005). "A monkey's tale: the origin of Plasmodium vivax as a human malaria parasite." Proc Natl Acad Sci U S A **102**(6): 1980-5.

Fernandez, V., Q. Chen, et al. (2002). "Mosaic-like transcription of var genes in single Plasmodium falciparum parasites." Mol Biochem Parasitol **121**(2): 195-203.

Fernandez, V., M. Hommel, et al. (1999). "Small, clonally variant antigens expressed on the surface of the Plasmodium falciparum-infected erythrocyte are encoded by the rif gene family and are the target of human immune responses." J Exp Med **190**(10): 1393-404.

Fernandez-Becerra, C., O. Pein, et al. (2005). "Variant proteins of Plasmodium vivax are not clonally expressed in natural infections." Mol Microbiol **58**(3): 648-58.

Feuerbach, F., V. Galy, et al. (2002). "Nuclear architecture and spatial positioning help establish transcriptional states of telomeres in yeast." Nat Cell Biol **4**(3): 214-21.

Figueiredo, L. and A. Scherf (2005). "Plasmodium telomeres and telomerase: the usual actors in an unusual scenario." Chromosome Res **13**(5): 517-24.

Figueiredo, L. M., L. H. Freitas-Junior, et al. (2002). "A central role for Plasmodium falciparum subtelomeric regions in spatial positioning and telomere length regulation." Embo J **21**(4): 815-24.

Figueiredo, L. M., L. A. Pirrit, et al. (2000). "Genomic organisation and chromatin structure of Plasmodium falciparum chromosome ends." Mol Biochem Parasitol **106**(1): 169-74.

Figueiredo, L. and A. Scherf (2005). "Plasmodium telomeres and telomerase: the usual actors in an unusual scenario." Chromosome Res **13**(5): 517-24. (a)

Figueiredo, L. M., E. P. Rocha, et al. (2005). "The unusually large Plasmodium telomerase reverse-transcriptase localizes in a discrete compartment associated with the nucleolus." Nucleic Acids Res **33**(3): 1111-22. (b)

Fischer, K., M. Chavchich, et al. (2003). "Ten families of variant genes encoded in subtelomeric regions of multiple chromosomes of Plasmodium chabaudi, a malaria species that undergoes antigenic variation in the laboratory mouse." Mol Microbiol **48**(5): 1209-23.

Flick, K. and Q. Chen (2004). "var genes, PfEMP1 and the human host." Mol Biochem Parasitol **134**(1): 3-9.

Florens, L., M. P. Washburn, et al. (2002). "A proteomic view of the Plasmodium falciparum life cycle." Nature **419**(6906): 520-6.

Fong, N. and D. L. Bentley (2001). "Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD." Genes Dev **15**(14): 1783-95.

Fourel, G., E. Revardel, et al. (1999). "Cohabitation of insulators and silencing elements in yeast subtelomeric regions." Embo J **18**(9): 2522-37.

Freitas-Junior, L. H., E. Bottius, et al. (2000). "Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum." Nature **407**(6807): 1018-22.

Freitas-Junior, L. H., R. Hernandez-Rivas, et al. (2005). "Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites." Cell **121**(1): 25-36.

Galinski, M. R. and V. Corredor (2004). "Variant antigen expression in malaria infections: posttranscriptional gene silencing, virulence and severe pathology." Mol Biochem Parasitol **134**(1): 17-25.

Gallup, J. L. and J. D. Sachs (2001). "The economic burden of malaria." Am J Trop Med Hyg **64**(1-2 Suppl): 85-96.

Gannoun-Zaki, L., A. Jost, et al. (2005). "A silenced Plasmodium falciparum var promoter can be activated in vivo through spontaneous deletion of a silencing element in the intron." Eukaryot Cell **4**(2): 490-2.

Gardner, P. P. and R. Giegerich (2004). "A comprehensive comparison of comparative RNA structure prediction approaches." BMC Bioinformatics **5**: 140.

Gardner, M. J., N. Hall, et al. (2002). "Genome sequence of the human malaria parasite Plasmodium falciparum." Nature **419**(6906): 498-511.

Gasser, S. M. and M. M. Cockell (2001). "The molecular biology of the SIR proteins." Gene **279**(1): 1-16.

Gilks, C. F., D. Walliker, et al. (1990). "Relationships between sequestration, antigenic variation and chronic parasitism in Plasmodium chabaudi chabaudi--a rodent malaria model." Parasite Immunol **12**(1): 45-64.

Gissot, M., P. Refour, et al. (2004). "Transcriptome of 3D7 and its gametocyte-less derivative F12 Plasmodium falciparum clones during erythrocytic development using a gene-specific microarray assigned to gene regulation, cell cycle and transcription factors." Gene **341**: 267-77.

Golightly, L. M., W. Mbacham, et al. (2000). "3' UTR elements enhance expression of Pgs28, an ookinete protein of Plasmodium gallinaceum." Mol Biochem Parasitol **105**(1): 61-70.

Gottschling, D. E., O. M. Aparicio, et al. (1990). "Position effect at S. cerevisiae telomeres: reversible repression of Pol II transcription." Cell **63**(4): 751-62.

Grewal, S. I. and D. Moazed (2003). "Heterochromatin and epigenetic control of gene expression." Science **301**(5634): 798-802.

Gruner, A. C., S. Hez-Deroubaix, et al. (2005). "Insights into the P. y. yoelii hepatic stage transcriptome reveal complex transcriptional patterns." Mol Biochem Parasitol **142**(2): 184-92.

Gruner, A. C., G. Snounou, et al. (2004). "The Py235 proteins: glimpses into the versatility of a malaria multigene family." Microbes Infect **6**(9): 864-73.

Hall, N., M. Karras, et al. (2005). "A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses." Science **307**(5706): 82-6.

Hall BG et al., 2004, Phylogenetic Trees Made Easy 2nd edition Sinauer Associates

Harding, H. P., I. Novoa, et al. (2000). "Regulated translation initiation controls stress-induced gene expression in mammalian cells." Mol Cell **6**(5): 1099-108.

Henikoff, S. (2000). "Heterochromatin function in complex genomes." Biochim Biophys Acta **1470**(1): O1-8.

Hiller, N. L., S. Bhattacharjee, et al. (2004). "A host-targeting signal in virulence proteins reveals a secretome in malarial infection." Science **306**(5703): 1934-7.

Holder, A. A., J. A. Guevara Patino, et al. (1999). "Merozoite surface protein 1, immune evasion, and vaccines against asexual blood stage malaria." Parassitologia **41**(1-3): 409-14.

Hommel, M., P. H. David, et al. (1983). "Surface alterations of erythrocytes in Plasmodium falciparum malaria. Antigenic variation, antigenic diversity, and the role of the spleen." J Exp Med **157**(4): 1137-48.

Hoopes, B. C., J. F. LeBlanc, et al. (1998). "Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II." J Mol Biol **277**(5): 1015-31.

Horrocks, P., R. Pinches, et al. (2004). "Variable var transition rates underlie antigenic variation in malaria." Proc Natl Acad Sci U S A **101**(30): 11129-34.

Horrocks, P., S. Kyes, et al. (2004). "Transcription of subtelomerically located var gene variant in Plasmodium falciparum appears to require the truncation of an adjacent var gene." Mol Biochem Parasitol **134**(2): 193-9. **(b)**

Irvine, R. A., I. G. Lin, et al. (2002). "DNA methylation has a local effect on transcription and histone acetylation." Mol Cell Biol **22**(19): 6689-96.

Janssen, C. S., M. P. Barrett, et al. (2002). "A large gene family for putative variant antigens shared by human and rodent malaria parasites." Proc Biol Sci **269**(1489): 431-6.

Janssen, C. S., R. S. Phillips, et al. (2004). "Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites." Nucleic Acids Res **32**(19): 5712-20.

John, C. C., A. M. Moormann, et al. (2005). "Correlation of high levels of antibodies to multiple pre-erythrocytic Plasmodium falciparum antigens and protection from infection." Am J Trop Med Hyg **73**(1): 222-8.

Jongwutiwes, S., C. Putaporntip, et al. (2005). "Mitochondrial genome sequences support ancient population expansion in Plasmodium vivax." Mol Biol Evol **22**(8): 1733-9.

Joshi-Barve, S., A. De Benedetti, et al. (1992). "Preferential translation of heat shock mRNAs in HeLa cells deficient in protein synthesis initiation factors eIF-4E and eIF-4 gamma." J Biol Chem **267**(29): 21038-43.

Joshi, M. B., D. T. Lin, et al. (1999). "Molecular cloning and nuclear localization of a histone deacetylase homologue in Plasmodium falciparum." Mol Biochem Parasitol **99**(1): 11-9.

Joy, D. A., X. Feng, et al. (2003). "Early origin and recent expansion of Plasmodium falciparum." Science **300**(5617): 318-21.

Jurica, M. S. and M. J. Moore (2003). "Pre-mRNA splicing: awash in a sea of proteins." Mol Cell **12**(1): 5-14.

Kalmykova, A. I., D. I. Nurminsky, et al. (2005). "Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster." Nucleic Acids Res **33**(5): 1435-44.

Kaviratne, M., S. M. Khan, et al. (2002). "Small variant STEVOR antigen is uniquely located within Maurer's clefts in Plasmodium falciparum-infected red blood cells." Eukaryot Cell **1**(6): 926-35.

Kyes, S. A., Z. Christodoulou, et al. (2003). "A well-conserved Plasmodium falciparum var gene shows an unusual stage-specific transcript pattern." Mol Microbiol **48**(5): 1339-48.

Khan, S. M., W. Jarra, et al. (2001). "Distribution and characterisation of the 235 kDa rhoptry multigene family within the genomes of virulent and avirulent lines of Plasmodium yoelii." Mol Biochem Parasitol **114**(2): 197-208.

Khan, S. M., W. Jarra, et al. (2001). "The 235 kDa rhoptry protein of Plasmodium (yoelii) yoelii: function at the junction." Mol Biochem Parasitol **117**(1): 1-10.

Knapp, B., U. Nau, et al. (1991). "Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of Plasmodium falciparum." J Biol Chem **266**(11): 7148-54.

Kozak, M. (2005). "A second look at cellular mRNA sequences said to function as internal ribosome entry sites." Nucleic Acids Res **33**(20): 6593-602.

Kraemer, S. M. and J. D. Smith (2003). "Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family." Mol Microbiol **50**(5): 1527-38.

Kumar, S., K. Tamura, et al. (2001). "MEGA2: molecular evolutionary genetics analysis software." Bioinformatics **17**(12): 1244-5.

Kwiatkowski, D. P. (2005). "How malaria has affected the human genome and what human genetics can teach us about malaria." Am J Hum Genet **77**(2): 171-92.

Kyes, S., P. Horrocks, et al. (2001). "Antigenic variation at the infected red cell surface in malaria." Annu Rev Microbiol **55**: 673-707.

Kyes, S. A., J. A. Rowe, et al. (1999). "Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum." Proc Natl Acad Sci U S A **96**(16): 9333-8.

Lavstsen, T., A. Salanti, et al. (2003). "Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions." Malar J **2**: 27.

Li, J., R. R. Gutell, et al. (1997). "Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite." J Mol Biol **269**(2): 203-13.

Li, Y., C. P. Darley, et al. (2002). "Plant expansins are a complex multigene family with an ancient evolutionary origin." Plant Physiol **128**(3): 854-64.

Li, Y. J., X. H. Fu, et al. (2004). "Opening the chromatin for transcription." Int J Biochem Cell Biol **36**(8): 1411-23.

Marti, M., R. T. Good, et al. (2004). "Targeting malaria virulence and remodeling proteins to the host erythrocyte." Science **306**(5703): 1930-3.

McLean, S. A., C. D. Pearson, et al. (1982). "Plasmodium chabaudi: antigenic variation during recrudescent parasitaemias in mice." Exp Parasitol **54**(3): 296-302.

McRobert, L., P. Preiser, et al. (2004). "Distinct trafficking and localization of STEVOR proteins in three stages of the Plasmodium falciparum life cycle." Infect Immun **72**(11): 6597-602.

Medica, D. L. and P. Sinnis (2005). "Quantitative dynamics of Plasmodium yoelii sporozoite transmission by infected anopheline mosquitoes." Infect Immun **73**(7): 4363-9.

Mendis, K., B. J. Sina, et al. (2001). "The neglected burden of Plasmodium vivax malaria." Am J Trop Med Hyg **64**(1-2 Suppl): 97-106.

Mignone, F., C. Gissi, et al. (2002). "Untranslated regions of mRNAs." Genome Biol **3**(3): REVIEWS0004.

Militello, K. T., M. Dodge, et al. (2004). "Identification of regulatory elements in the Plasmodium falciparum genome." Mol Biochem Parasitol **134**(1): 75-88.

Moazed, D. (2001). "Common themes in mechanisms of gene silencing." Mol Cell **8**(3): 489-98.

Newbold, C., P. Warn, et al. (1997). "Receptor-specific adhesion and clinical disease in Plasmodium falciparum." Am J Trop Med Hyg **57**(4): 389-98.

Newbold, C. I., R. Pinches, et al. (1992). "Plasmodium falciparum: the human agglutinating antibody response to the infected red cell surface is predominantly variant specific." Exp Parasitol **75**(3): 281-92.

Niimura, Y. and M. Nei (2005). "Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods." Proc Natl Acad Sci U S A **102**(17): 6039-44.

Noviyanti, R., G. V. Brown, et al. (2001). "Multiple var gene transcripts are expressed in Plasmodium falciparum infected erythrocytes selected for adhesion." Mol Biochem Parasitol **114**(2): 227-37.

O'Donnell, R. A., L. H. Freitas-Junior, et al. (2002). "A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of Plasmodium falciparum chromosomes." Embo J **21**(5): 1231-9.

Osta, M., L. Gannoun-Zaki, et al. (2002). "A 24 bp cis-acting element essential for the transcriptional activity of Plasmodium falciparum CDP-diacylglycerol synthase gene promoter." Mol Biochem Parasitol **121**(1): 87-98.

Owen, C. A., K. A. Sinha, et al. (1999). "Chromosomal organisation of a gene family encoding rhoptry proteins in Plasmodium yoelii." Mol Biochem Parasitol **99**(2): 183-92.

Pace, T., C. Birago, et al. (1998). "Developmental regulation of a Plasmodium gene involves the generation of stage-specific 5' untranslated sequences." Mol Biochem Parasitol **97**(1-2): 45-53.

Perkins, S. L. and J. J. Schall (2002). "A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences." J Parasitol **88**(5): 972-8.

Pesole, G., F. Mignone, et al. (2001). "Structural and functional features of eukaryotic mRNA untranslated regions." Gene **276**(1-2): 73-81.

Pollard, A. J., A. R. Krainer, et al. (2002). "Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3'-splice Site." J Biol Chem **277**(18): 15241-51.

Polson, H. E. and M. J. Blackman (2005). "A role for poly(dA)poly(dT) tracts in directing activity of the Plasmodium falciparum calmodulin gene promoter." Mol Biochem Parasitol **141**(2): 179-89.

Ponnudurai, T., A. H. Lensen, et al. (1991). "Feeding behaviour and sporozoite ejection by infected Anopheles stephensi." Trans R Soc Trop Med Hyg **85**(2): 175-80.

Preiser, P. R. and W. Jarra (1998). "Plasmodium yoelii: differences in the transcription of the 235-kDa rhoptry protein multigene family in lethal and nonlethal lines." Exp Parasitol **89**(1): 50-7.

Preiser, P. R., W. Jarra, et al. (1999). "A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria." Nature **398**(6728): 618-22.

Preiser, P. R., S. Khan, et al. (2002). "Stage-specific transcription of distinct repertoires of a multigene family during Plasmodium life cycle." Science **295**(5553): 342-5.

Preiser, P., L. Renia, et al. (2004). "Antibodies against MAEBL ligand domains M1 and M2 inhibit sporozoite development in vitro." Infect Immun **72**(6): 3604-8.

Przyborski, J. M. and M. Lanzer (2005). "Protein transport and trafficking in Plasmodium falciparum-infected erythrocytes." Parasitology **130**(Pt 4): 373-88.

Ralph, S. A., C. Scheidig-Benatar, et al. (2005). "Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations." Proc Natl Acad Sci U S A **102**(15): 5414-9.

Rich, S. M., M. C. Licht, et al. (1998). "Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum." Proc Natl Acad Sci U S A **95**(8): 4425-30.

Roberts, D. J., A. G. Craig, et al. (1992). "Rapid switching to multiple antigenic and adhesive phenotypes in malaria." Nature **357**(6380): 689-92.

Rosenberg, R. and J. Rungsiwongse (1991). "The number of sporozoites produced by individual malaria oocysts." Am J Trop Med Hyg **45**(5): 574-7.

Rosenberg, R., R. A. Wirtz, et al. (1990). "An estimation of the number of malaria sporozoites ejected by a feeding mosquito." Trans R Soc Trop Med Hyg **84**(2): 209-12.

Rusche, L. N., A. L. Kirchmaier, et al. (2003). "The establishment, inheritance, and function of silenced chromatin in Saccharomyces cerevisiae." Annu Rev Biochem **72**: 481-516.

Ruvalcaba-Salazar, O. K., M. del Carmen Ramirez-Estudillo, et al. (2005). "Recombinant and native Plasmodium falciparum TATA-binding-protein binds to a specific TATA box element in promoter regions." Mol Biochem Parasitol **140**(2): 183-96.

Sacci, J. B., Jr., J. M. Ribeiro, et al. (2005). "Transcriptional analysis of in vivo Plasmodium yoelii liver stage gene expression." Mol Biochem Parasitol **142**(2): 177-83.

Sam-Yellowe, T. Y., L. Florens, et al. (2004). "A Plasmodium gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling." Genome Res **14**(6): 1052-9.

Sandell, L. L. and V. A. Zakian (1992). "Telomeric position effect in yeast." Trends Cell Biol **2**(1): 10-4.

Scherf, A., L. M. Figueiredo, et al. (2001). "Plasmodium telomeres: a pathogen's perspective." Curr Opin Microbiol **4**(4): 409-14.

Scherf, A., R. Hernandez-Rivas, et al. (1998). "Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum." Embo J **17**(18): 5418-26.

Shankaranarayana, G. D., M. R. Motamedi, et al. (2003). "Sir2 regulates histone H3 lysine 9 methylation and heterochromatin assembly in fission yeast." Curr Biol **13**(14): 1240-6.

Sherman, I. W., S. Eda, et al. (2004). "Erythrocyte aging and malaria." Cell Mol Biol (Noisy-le-grand) **50**(2): 159-69.

Shi, Q., A. Cernetich, et al. (2005). "Alteration in host cell tropism limits the efficacy of immunization with a surface protein of malaria merozoites." Infect Immun **73**(10): 6363-71.

Singh, N., P. Preiser, et al. (2004). "Conservation and developmental control of alternative splicing in maebl among malaria parasites." J Mol Biol **343**(3): 589-99.

Smith, J. D., C. E. Chitnis, et al. (1995). "Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes." Cell **82**(1): 101-10.

Shinkai, Y., G. Rathbun, et al. (1992). "RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement." Cell **68**(5): 855-67.

Snounou, G., W. Jarra, et al. (2000). "Malaria multigene families: the price of chronicity." Parasitol Today **16**(1): 28-30.

Spielmann, T., D. J. Fergusen, et al. (2003). "etramps, a new Plasmodium falciparum gene family coding for developmentally regulated and highly charged membrane proteins located at the parasite-host cell interface." Mol Biol Cell **14**(4): 1529-44.

Sriwilaijareon, N., S. Petmitr, et al. (2002). "Stage specificity of Plasmodium falciparum telomerase and its inhibition by berberine." Parasitol Int **51**(1): 99-103.

Sullivan, D. J., Jr., Y. M. Ayala, et al. (1996). "An unexpected 5' untranslated intron in the P. falciparum genes for histidine-rich proteins II and III." Mol Biochem Parasitol **83**(2): 247-51.

Sutherland, C. J. (2001). "Stevor transcripts from Plasmodium falciparum gametocytes encode truncated polypeptides." Mol Biochem Parasitol **113**(2): 331-5.

Taylor, H. M., S. A. Kyes, et al. (2000). "A study of var gene transcription in vitro using universal var gene primers." Mol Biochem Parasitol **105**(1): 13-23.

Tham, W. H., J. S. Wyithe, et al. (2001). "Localization of yeast telomeres to the nuclear periphery is separable from transcriptional repression and telomere stability functions." Mol Cell **8**(1): 189-99.

van Noort, V. and M. A. Huynen (2005). "Combinatorial gene regulation in Plasmodium falciparum." Trends Genet.

van Spaendonk, R. M., J. Ramesar, et al. (2001). "Functional equivalence of structurally distinct ribosomes in the malaria parasite, Plasmodium berghei." J Biol Chem **276**(25): 22638-47.

Vazquez-Macias, A., P. Martinez-Cruz, et al. (2002). "A distinct 5' flanking var gene region regulates Plasmodium falciparum variant erythrocyte surface antigen expression in placental malaria." Mol Microbiol **45**(1): 155-67.

Voss, T. S., M. Kaestli, et al. (2003). "Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters." Mol Microbiol **48**(6): 1593-607.

Voss, T. S., J. K. Thompson, et al. (2000). "Genomic distribution and functional characterisation of two distinct and conserved Plasmodium falciparum var gene 5' flanking sequences." Mol Biochem Parasitol **107**(1): 103-15.

Watanabe, J., M. Sasaki, et al. (2002). "Analysis of transcriptomes of human malaria parasite Plasmodium falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs." Gene **291**(1-2): 105-13.

Weiss, K. and R. T. Simpson (1998). "High-resolution structural analysis of chromatin at specific loci: Saccharomyces cerevisiae silent mating type locus HMLalpha." Mol Cell Biol **18**(9): 5392-403.

Wickham, M. E., J. K. Thompson, et al. (2003). "Characterisation of the merozoite surface protein-2 promoter using stable and transient transfection in Plasmodium falciparum." Mol Biochem Parasitol **129**(2): 147-56.

Winter, G., S. Kawai, et al. (2005). "SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes." J Exp Med **201**(11): 1853-63.

Yiu, G. K., W. Gu, et al. (1994). "Heterogeneity in the 5' untranslated region of mouse cytochrome cT mRNAs leads to altered translational status of the mRNAs." Nucleic Acids Res **22**(22): 4599-606.

Zardoya, R., E. Abouheif, et al. (1996). "Evolutionary analyses of hedgehog and Hoxd-10 genes in fish species closely related to the zebrafish." Proc Natl Acad Sci U S A **93**(23): 13036-41.

Zhou, Z., L. J. Licklider, et al. (2002). "Comprehensive proteomic analysis of the human spliceosome." Nature **419**(6903): 182-5.