



This work is protected by copyright and other intellectual property rights and duplication or sale of all or part is not permitted, except that material may be duplicated by you for research, private study, criticism/review or educational purposes. Electronic or print copies are for your own personal, non-commercial use and shall not be passed to any other individual. No quotation may be published without proper acknowledgement. For any other use, or to quote extensively from the work, permission must be obtained from the copyright holder/s.

This electronic version of the thesis has been edited solely to ensure compliance with copyright legislation and excluded material is referenced in the text. The full, final, examined and awarded version of the thesis is available for consultation in hard copy via the University Library.

A spatial and temporal analysis
of *Plasmodium falciparum*
transcription

Karen Russell

PhD Thesis

December, 2014

Keele University

STUDENT DECLARATION

ABSTRACT

Developmentally-linked gene expression is critical to the success of the human malaria parasite *Plasmodium falciparum* in ensuring colonisation, adaptation, replication and transmission during its complex life cycle as well as the manifestation of disease in humans. Yet, despite the wealth of high-throughput transcriptomic data, our understanding of the organisation of the transcriptional unit outside of the open reading frame is limited.

The objectives of this study were directed towards understanding how intergenic space is organised over the entire *P. falciparum* genome and determining the likely spatio-temporal organisation of transcripts over these intergenic regions. In addition, as homopolymeric poly dA.dT are significantly overrepresented within these regions, a spatial analysis of poly dA.dT tract positional bias was undertaken and correlated with available nucleosome positioning data. These studies in *P. falciparum* were supported with comparative analyses using a range of other Apicomplexan parasites. Finally, the role of the 5' untranslated region in directing transcriptional and translational efficiency for a typical housekeeping gene was investigated. Towards these aims a range of approaches were employed including bioinformatics, comparative genomics, data modelling and reporter gene assays.

The findings presented in this thesis extend our understanding of the transcriptional landscape in this important human pathogen, generating models that can be experimentally validated when new RNAseq datasets become available. Ideas relating to how different selective forces are at play in shaping the organisation and sequence of intergenic regions are also presented. Moreover, we demonstrate comparable organisations of intergenic regions and homopolymer tracts within a number of Apicomplexan parasites, many important for human and animal health, providing the basis for a comparative approach to understanding transcriptional processes across this medically important phylum.

CONTENTS

CHAPTER 1 INTRODUCTION	1
1.1 MALARIA.....	1
1.2 THE LIFE CYCLE OF <i>P. FALCIPARUM</i>	5
1.3 UNDERSTANDING THE CONTROL OF GENE EXPRESSION – MOVING INTO THE POST-GENOMIC ERA	10
1.3.1 <i>Understanding the control of gene expression- the impact of Transcriptomics</i>	14
1.3.2 <i>Understanding the control of gene expression- the impact of post-transcriptional and post-translational mechanisms</i>	18
1.3.3 <i>Understanding the control of gene expression- the impact of promoter-based cis-trans interactions</i>	22
1.3.4 <i>Understanding the control of gene expression – the impact of Chromatin organization and epigenetic regulators</i>	26
1.4 CONTROL OF GENE EXPRESSION IN <i>P. FALCIPARUM</i> IS MEDIATED AT ALL LEVELS OF TRANSCRIPTION AND TRANSLATION – AND RELIES ON SIGNALS IN THE INTERGENIC REGIONS.....	34
1.5 AIMS OF MY RESEARCH.....	39
CHAPTER 2 MATERIALS AND METHODS	41
2.1 LABORATORY STOCK REAGENTS	41
2.2 CULTURE OF <i>PLASMODIUM FALCIPARUM</i>	44
2.2.1 <i>P. falciparum clones</i>	44
2.2.2 <i>P. falciparum cell culture medium</i>	44
2.2.3 <i>Human Red Blood Cell preparation</i>	44
2.2.4 <i>Maintenance of P. falciparum by continuous cell culture</i>	45
2.2.5 <i>Culture Synchronisation</i>	45
2.2.6 <i>Transfection</i>	46
2.3 DNA METHODS	46
2.3.1 <i>P. falciparum gDNA isolation</i>	46

2.3.2 Minipreparation of Plasmid DNA.....	47
2.3.3 Maxipreparation of Plasmid Dna.....	48
2.3.4 Spectroscopic analysis of DNA.....	48
2.3.5 Polymerase Chain Reaction (PCR).....	49
2.3.6 Gel Electrophoresis and Imaging.....	49
2.3.7 TOPO cloning.....	50
2.3.8 Restriction digest.....	51
2.3.9 Quantitative PCR.....	51
2.3.10 Sequencing.....	51
2.4 RNA METHODS.....	52
2.4.1 <i>P. falciparum</i> RNA extraction.....	52
2.4.2 Northern Blots.....	53
2.4.3 5' Rapid Amplification of cDNA ends (RACE).....	54
2.4.4 3' RACE.....	56
2.5 LUCIFERASE ASSAY.....	57
2.6 EXPERIMENTAL OLIGONUCLEOTIDES.....	58
2.7 BIOINFORMATICS.....	62
2.7.1 Intergenic Distance Analyses.....	62
2.7.2 Gene Selection for Northern Blots.....	67
2.7.3 Transcript Apportionment.....	68
2.7.4 Poly.....	70
2.7.5 Organisation of homopolymeric dA.dT tracts around the ORF.....	72

CHAPTER 3 ORGANIZATION OF INTERGENIC SPACE IN *P. FALCIPARUM* AND OTHER

APICOMPLEXAN PARASITES.....	74
3.1 INTRODUCTION.....	74
3.2 CAPTURE AND CHARACTERIZATION OF INTERGENIC DATASETS.....	75
3.3 ANALYSIS OF INTERGENIC SPACE IN <i>P. FALCIPARUM</i>	77

3.4 INTERGENIC DISTANCES ARE LONGER IN <i>P. FALCIPARUM</i> SUBTELOMERIC REGIONS	82
3.5 THE LEVEL OF TRANSCRIPTIONAL ACTIVITY DOES NOT APPEAR TO CORRELATE WITH THE SIZE OF THE FLANKING IGR	86
3.6 DIFFERENTIAL APPORTIONMENT OF THE INTERGENIC SPACE IS A FEATURE OF GENOME ORGANISATION IN A RANGE OF APICOMPLEXAN PARASITES.....	89
3.6.1 <i>Analysis of the ratio of intergenic space organization</i>	96
3.7 DISCUSSION	100
CHAPTER 4 A COMPARATIVE ANALYSIS OF UTR SIZE IN <i>P. FALCIPARUM</i>	104
4.1 INTRODUCTION.....	104
4.2 EXPLORING THE RELATIONSHIP BETWEEN ORF AND TRANSCRIPT SIZE	106
4.2.1 <i>Collection of Northern Blot data</i>	106
4.2.2 <i>Analysis of UTR sizes derived from Northern Blot data</i>	113
4.3 ANALYSIS OF UTR SIZES DERIVED FROM EST DATA	117
4.4 DISCUSSION.....	124
CHAPTER 5 MODELLING SPATIAL ORGANIZATION OF TRANSCRIPTS OVER INTERGENIC SPACE IN <i>P. FALCIPARUM</i>	127
5.1 INTRODUCTION.....	127
5.2 MODELLING APPROACH.....	130
5.3 RESULTS.....	136
5.3.1 <i>Testing of Scenario A</i>	136
5.3.2 <i>Scenario A</i>	137
5.3.3 <i>Scenario B</i>	140
5.4 GENES IN A HEAD TO HEAD ORIENTATION WITH THE SAME TEMPORAL PROFILE HAVE SMALLER IGRS	142
5.5 DISCUSSION	145
CHAPTER 6 ORGANIZATION OF HOMOPOLYMERIC DA.DT TRACTS IN <i>P. FALCIPARUM</i> AND OTHER APICOMPLEXAN PARASITES	149
6.1 INTRODUCTION.....	149

6.2 COMPARATIVE ANALYSIS OF HOMOPOLYMER TRACT FREQUENCY AND LENGTH IN THE PROXIMAL INTERGENIC REGIONS OF APICOMPLEXAN PARASITES	153
6.3 SPATIAL ANALYSIS OF POLY DA.DT TRACTS IN PROXIMAL FLANKING INTERGENIC SEQUENCES.....	171
6.4 SPATIAL ANALYSIS OF POLY DA.DT TRACTS OVER THE CORE PROMOTER IN <i>P. FALCIPARUM</i>	179
6.5 DISCUSSION	183
CHAPTER 7 FUNCTIONAL ANALYSIS OF THE 5' UTR IN <i>P. FALCIPARUM</i>.....	189
7.1 INTRODUCTION.....	189
7.2 SELECTION AND CHARACTERIZATION OF PFD0660w AS CANDIDATE GENE FOR THIS STUDY	190
7.3 GENERATION OF REPORTER CONSTRUCTS FOR THIS STUDY	192
7.4 TRANSFECTION OF FL, $\Delta 1$, $\Delta 2$ AND $\Delta 3$	199
7.5 ANALYSIS OF LUCIFERASE EXPRESSION IN THE FL TRANSFECTANT.....	203
7.6 ANALYSIS OF ABSOLUTE AND TEMPORAL EXPRESSION IN THE DELETION CONSTRUCT SERIES.....	206
7.7 DISCUSSION	209
CHAPTER 8 CONCLUSIONS.....	214
8.1 INTERGENIC REGIONS AND THE TRANSCRIPTIONAL LANDSCAPE OF <i>P. FALCIPARUM</i>	214
8.2 WHAT IS THE FUNCTION OF POLY DA.DT TRACTS IN <i>PLASMODIUM SPP.</i> ?.....	221
8.3 PLASTICITY OF THE UTR.....	223
APPENDICES	226
APPENDIX A - HORROCKS <i>ET AL.</i> , 2009	226
APPENDIX B - HASENKAMP <i>ET AL.</i> , 2013	228
APPENDIX C - RUSSELL <i>ET AL.</i> , 2013.....	230
APPENDIX D – SUPPLEMENTARY FILES CHAPTER 4.....	232
<i>D-1 Reference list for Northern Blot data taken from the literature</i>	<i>232</i>
<i>D-2 Microarray data providing peak transcript data for the 105 genes in the Northern Blot cohort ...</i>	<i>237</i>
APPENDIX E – SUPPLEMENTARY FILES CHAPTER 7	239
<i>Appendix E-1 ORF adjacent homopolymeric dA and dT tracts</i>	<i>239</i>

<i>Appendix E-2 Cloning primer locations</i>	240
<i>Appendix E-3 $\Delta 4$ plasmid for transfection and proof of integration</i>	242
<i>Appendix E-4 RT-PCR</i>	243
<i>Appendix E-5 Plot representing %GC over the 5' intergenic region of PFD0660w and predicted AP2 binding sites</i>	244
APPENDIX F – SUPPLEMENTARY FILES CHAPTER 8	245
<i>F-1 FIRE ANALYSIS GROUP A</i>	245
<i>F-2 INTERACTIONS AMONGST GROUP A MOTIFS</i>	246
<i>F-3 FIRE ANALYSIS GROUP B</i>	247
<i>F-4 INTERACTIONS AMONGST GROUP B MOTIFS</i>	248
<i>F-5 FIRE ANALYSIS GROUP C</i>	249
<i>F-6 INTERACTIONS AMONGST GROUP C MOTIFS</i>	250
<i>F-7 FIRE ANALYSIS AND INTERACTION AMONGST MOTIFS GROUP D</i>	251
<i>F-8 FIRE MOTIF COMPARISON BETWEEN ALL DATASETS</i>	252
APPENDIX G - HASENKAMP <i>ET AL</i> , 2012	254
APPENDIX H - RUSSELL <i>ET AL</i> , 2014	256
REFERENCES	257

TABLES AND FIGURES

FIGURE 1-1 (A) GLOBAL DISTRIBUTION OF <i>P. FALCIPARUM</i> AND <i>P. VIVAX</i> IN 2012 AND (B) WORLDWIDE MALARIA INPATIENT DEATHS RECORDED IN 2012.	2
FIGURE 1-2 THE LIFE-CYCLE OF <i>P. FALCIPARUM</i>	6
FIGURE 1-3 IDC TRANSCRIPTOME PHASEOGRAM.....	13
FIGURE 1-4 IDC TRANSCRIPTOME AND PROTEOME PHASEOGRAM.	21
TABLE 1-1 FUNCTIONAL PROMOTER ASSAYS IN <i>P. FALCIPARUM</i>	23
FIGURE 1-5 LIST OF AP2 PROTEINS IN <i>P. FALCIPARUM</i>	26
FIGURE 1-6 CHROMATIN DOMAINS OF THE <i>P. FALCIPARUM</i> GENOME.....	28
FIGURE 1-7 OVERVIEW OF REGULATION OF GENE EXPRESSION IN <i>P. FALCIPARUM</i>	36
TABLE 2-1 OLIGONUCLEOTIDES FOR NORTHERN BLOT ANALYSES.....	58
TABLE 2-2 OLIGONUCLEOTIDES FOR PFD0660w 5' UTR DELETION STUDY	60
TABLE 2-3 OLIGONUCLEOTIDES FOR GENOMIC INTEGRATION INTO <i>ATTB</i> LOCUS.....	61
TABLE 2-4 OLIGONUCLEOTIDES FOR QPCR OF LUCIFERASE COPY NUMBER AND EXPRESSION.....	61
TABLE 2-5 OLIGONUCLEOTIDES FOR 5' AND 3' RACE	61
FIGURE 2-1 SCHEMATIC REPRESENTING WORKFLOW TO CAPTURE INTERGENIC DISTANCES FOR HH, HT, TH AND TT GENE FLANKING REGIONS	62
FIGURE 2-2 SCHEMATIC REPRESENTING WORKFLOW TO SECURE INTERGENIC DISTANCES IN DIFFERENT <i>P. FALCIPARUM</i> CHROMOSOME COMPARTMENTS.	65
FIGURE 2-3 SCHEMATIC REPRESENTING WORK-FLOW TO CATEGORISE TEMPORAL WINDOWS OF TRANSCRIPTION DURING INTRAERYTHROCYTIC SCHIZOGONY FOR GENE-FLANKING PAIRS INTERGENIC REGIONS.	65
FIGURE 2-4 SCHEMATIC REPRESENTING THE WORKFLOW USED TO MODEL TRANSCRIPT APPORTIONMENT IN <i>P. FALCIPARUM</i>	68
FIGURE 2-5 SCHEMATIC REPRESENTING THE WORKFLOW FOR USE OF POLY.	70
FIGURE 2-6 SCHEMATIC ILLUSTRATING THE WORKFLOW FOR THE USE OF MOTIF.	72
FIGURE 3-1 OVERVIEW OF THE EXTRACTION OF RELEVANT INFORMATION FROM A .GFF FILE AND ALLOCATION OF THE INTERGENIC DISTANCE TO THE RELEVANT GENE ORIENTATION FILE	76

FIGURE 3-2 FREQUENCY HISTOGRAM REPRESENTATION OF <i>P. FALCIPARUM</i> INTERGENIC DISTANCE DATA WHEN SUB-GROUPED BY FLANKING GENE ORIENTATION	77
TABLE 3-1 RANGE OF VALUES WITHIN EACH INTERGENIC DISTANCE DATASET	78
TABLE 3-2 MEAN AND MEDIAN VALUES FOR EACH INTERGENIC DISTANCE DATASET	78
FIGURE 3-3 <i>P. FALCIPARUM</i> INTERGENIC DATA GROUPED UPON FLANKING GENE ORIENTATION	79
FIGURE 3-4 FILE CONCATENATION AND RE-TERMING OF DATASETS.....	80
TABLE 3-3 RATIONALISATION OF INTERGENIC DATASETS.....	81
FIGURE 3-5 FREQUENCY HISTOGRAM OF THE <i>P. FALCIPARUM</i> IGR TYPES A,B,C AND BOXPLOTS SHOWING 100% AND 2.5-97.5% OF THE DATA RESPECTIVELY	81
TABLE 3-4 CHROMOSOMAL BREAKPOINTS DEEMED TO BE CHROMOSOME INTERNAL AND SUBTELOMERIC.	84
FIGURE 3-6 COMPARISON OF THE SIZE OF CHROMOSOMAL INTERNAL IGRS WITH SUBTELOMERIC IGRS.....	85
FIGURE 3-7 CORRELATION BETWEEN THE SIZE OF THE IGR AND MRNA ABUNDANCE (MRNAA)	88
TABLE 3-5 QUARTILE VALUES OF IGR LENGTH AND MRNAA.....	89
TABLE 3-6 RAW DATA DOWNLOAD INFORMATION FOR THE THIRTEEN ORGANISMS INVESTIGATED	91
TABLE 3-7 COMPARISON OF HH, HT, TH AND TT INTERGENIC GROUP MEDIANS FOR THE THIRTEEN ORGANISMS INVESTIGATED	92
FIGURE 3-8 HISTOGRAM REPRESENTATION OF THE A, B AND C DATASETS FOR ALL ORGANISMS.....	94
FIGURE 3-9 BOX AND WHISKERS PLOTS FOR THE 95% INTERGENIC DISTANCE ORIENTATION GROUPS, A, B AND C FOR ALL ORGANISMS INVESTIGATED	96
TABLE 3-8 AT-CONTENT AND THE IGR SIZE AND COUNT FOR EACH A, B AND C DATASET FOR EACH OF THE 13 ORGANISMS INVESTIGATED	98
FIGURE 3-10 COMPARISON OF AT CONTENT AND GENE DENSITY AGAINST MEDIAN IGR SIZE (BP)	99
TABLE 3-9 COMPARISON OF THE SIZE AND ORGANISATION OF IGR FROM ORGANISMS USED IN THIS STUDY	99
FIGURE 4-1 EXAMPLE NORTHERN BLOT DATA GENERATED <i>DE NOVO</i> FOR THIS STUDY	108
TABLE 4-1 COHORT OF 105 ORF FROM <i>P. FALCIPARUM</i> FOR WHICH NORTHERN BLOT DATA WAS COLLATED	108
FIGURE 4-2 UTR LENGTHS FROM NORTHERN BLOT DATA	114
FIGURE 4-3 LINEAR REGRESSION ANALYSIS OF THE TRANSCRIPT AND ORF SIZE.....	115
TABLE 4-2 COMPARATIVE DATA BETWEEN UTR SIZES DETERMINED FROM NORTHERN BLOT AND EST SOURCES.....	117
FIGURE 4-4 COMPARATIVE ANALYSES OF ORF FOR WHICH NORTHERN BLOT AND EST DATA WERE AVAILABLE.....	122

FIGURE 5-1 SCENARIO A	132
FIGURE 5-2 SCENARIO B	132
FIGURE 5-3 INPUT DATA FOR AND METHODOLOGY AND DECISION MAKING PROCESSES OF INTERGENIC.DIST.2CLASH.PL AND INTERGENIC.DIST.2CLASH.INCL.FLANKING.PL	134
FIGURE 5-4 EXAMPLE OUTPUT DATA FOR INTERGENIC.DIST.2CLASH.PL.....	135
FIGURE 5-5 INTERGENIC.DIST.2CLASH.PL OUTPUT DATA FOR THE 105 NORTHERN BLOT GENE COHORT	137
FIGURE 5-6 PLOT OF MEAN % FAIL RATE AGAINST % UTR APPORTIONED TO THE 5' FOR THE WHOLE <i>P. FALCIPARUM</i> GENOME USING PREDICTED UTR SIZES OF 600 TO 1800BP (200BP INCREMENTS) USING SCENARIO A.....	138
FIGURE 5-7 VISUAL REPRESENTATIONS DEPICTING THE MEDIAN IGR SIZE DEPENDENT UPON THE FLANKING GENE ORIENTATION, AND THE ORIENTATION OF THE THREE GENES WITHIN EACH TRIPLET (FORWARD AND REVERSE) THAT ARE CONTAINED WITHIN EACH GROUP.	139
FIGURE 5-8 GRANULATION OF INTERGENIC.DIST.2CLASH.PL OUTPUT DATA	140
FIGURE 5-9 PLOT OF MEAN % FAIL RATE AGAINST % UTR APPORTIONED TO THE 5' FOR THE WHOLE <i>P. FALCIPARUM</i> GENOME USING PREDICTED UTR SIZES OF 600 TO 1800BP IN 200BP INCREMENTS USING INTERGENIC.DIST.2CLASH.INCL.FLANKING.PL (SCENARIO B).	141
TABLE 5-1 COMPARISON OF INTERGENIC REGION SIZE FOR WHOLE <i>P. FALCIPARUM</i> GENOME AND CO-LOCATED CO-TRANSCRIBED GENES ONLY	144
FIGURE 5-10 ANALYSIS OF TEMPORAL CO-TRANSCRIPTION USING <i>P. FALCIPARUM</i> INTERGENIC REGION SIZE.....	144
FIGURE 6-1 PHYLOGENETIC RELATIONSHIP BETWEEN APICOMPLEXAN PARASITES USED IN THIS STUDY	154
TABLE 6-1 GENOMIC AT CONTENT, GENE DENSITY AND IGR WINDOW SIZE FOR EACH APICOMPLEXAN ORGANISM EVALUATED ...	155
FIGURE 6-2 EXAMPLE OUTPUT FROM POLY ANALYSIS OF HOMOPOLYMER TRACT FREQUENCY	157
FIGURE 6-3 REPRESENTATION OF HOMOPOLYMERIC TRACTS IN THE PROXIMAL UPSTREAM AND DOWNSTREAM INTERGENIC FLANKING REGIONS OF <i>PLASMODIUM SPP.</i> AND <i>CRYPTOSPORIDIUM SPP.</i>	158
FIGURE 6-4 REPRESENTATION OF HOMOPOLYMERIC TRACTS IN THE PROXIMAL UPSTREAM AND DOWNSTREAM INTERGENIC FLANKING REGIONS OF THE COCCIDIAN AND PIROPLASMIDA ORGANISMS.....	161
FIGURE 6-5 REPRESENTATION OF HOMOPOLYMERIC TRACTS IN CODING SEQUENCE OF THE COCCIDIAN AND PIROPLASMIDA ORGANISMS.....	163

TABLE 6-2 UPSTREAM AND DOWNSTREAM VALUES FOR FRACTION _i , N _{MAX} OBS, N _{MAX} EXP, P, R AND SLOPE _R FOR ALL APICOMPLEXAN ORGANISMS INVESTIGATED	166
FIGURE 6-6 COMPARATIVE ANALYSIS OF OVERREPRESENTATION OF HOMOPOLYMERIC TRACTS AS A FUNCTION OF NUCLEOTIDE CONTENT	167
FIGURE 6-7 COMPARATIVE ANALYSIS OF OVERPROPORTIONMENT OF HOMOPOLYMERIC TRACTS AS A FUNCTION OF NUCLEOTIDE CONTENT	169
FIGURE 6-8 COMPARATIVE ANALYSIS OF OVERPROPORTIONMENT OF HOMOPOLYMERIC TRACTS AS A FUNCTION OF SIZE OF INTERGENIC REGION (IGR)	170
FIGURE 6-9 SPATIAL DISTRIBUTION OF POLY DA.DT TRACTS IN PROXIMAL FLANKING INTERGENIC REGIONS OF THREE <i>PLASMODIUM SPP.</i>	173
FIGURE 6-10 SPATIAL DISTRIBUTION OF POLY DA.DT TRACTS OVER TRANSLATIONAL START AND STOP SITES OF THREE <i>PLASMODIUM SPP.</i>	174
FIGURE 6-11 SPATIAL DISTRIBUTION OF POLY DA.DT TRACTS OVER TRANSLATIONAL START AND STOP SITES OF THREE <i>PLASMODIUM SPP.</i> (2)	175
FIGURE 6-12 POLY DA.DT ENRICHMENT PROXIMAL TO THE TRANSLATIONAL START AND STOP SITE OF <i>P. YOELII</i> , <i>P. BERGHEI</i> , <i>CRYPTOSPORIDIUM SPP.</i> , <i>THEILERIA SPP.</i> , <i>T. GONDII</i> AND <i>N. CANINUM</i>	177
FIGURE 6-13 SPATIAL DISTRIBUTION OF NUCLEOSOME OCCUPANCY AND POLY DA.DT TRACTS OVER TRANSLATIONAL START AND STOP SITES OF <i>P. FALCIPARUM</i>	179
FIGURE 6-14 SPATIAL DISTRIBUTION OF NUCLEOSOME OCCUPANCY AND POLY DA.DT TRACTS OVER <i>P. FALCIPARUM</i> PREDICTED CORE PROMOTERS OF VARYING CONFIDENCE	181
FIGURE 6-15 SPATIAL DISTRIBUTION OF NUCLEOSOME OCCUPANCY AND POLY DA.DT TRACTS OVER PREDICTED CORE PROMOTERS IN <i>P. FALCIPARUM</i>	182
FIGURE 7-1 PFD0660W OVERVIEW AND EXPRESSION PROFILE	192
FIGURE 7-2 SCHEMATIC REPRESENTING THE SUB-CLONING STRATEGY EMPLOYED HERE	195
FIGURE 7-3 PRIMER LOCATIONS OVER <i>LUC</i> AND PFD0660W ORF	196
FIGURE 7-4 CONFIRMATION OF REPORTER PLASMIDS BY RESTRICTION DIGEST	197
FIGURE 7-5 RESTRICTION PLASMID MAPS	197

FIGURE 7-6 SCHEMATIC DEPICTION OF THE POSITION OF UPSTREAM FLANKING SEQUENCES IN FL, $\Delta 1$, $\Delta 2$ AND $\Delta 3$ CONSTRUCTS USED IN THIS STUDY.	199
FIGURE 7-7 SCHEMATIC REPRESENTATION OF INTEGRASE-MEDIATED INSERTION OF THE LUCIFERASE REPORTER CASSETTE INTO THE <i>cg6</i> LOCUS ON CHROMOSOME 7	200
FIGURE 7-8 CONFIRMATION OF INTEGRATION AND QPCR.....	202
FIGURE 7-9 STAGE-SPECIFIC NORTHERN BLOT ANALYSIS OF THE FL CONSTRUCT	203
FIGURE 7-10 PREDICTED UTR FROM DBEST	205
FIGURE 7-11 NORTHERN BLOTS OF THE TRANSFECTION SERIES.....	207
FIGURE 7-12 STAGE-SPECIFIC LUCIFERASE ASSAY.....	208
FIGURE 8-1 FIRE MOTIFS COMMON TO ALL DATASETS AND MOTIF INTERACTION MAP	217
FIGURE 8-2 MOTIF HEATMAP OF EXPRESSION THROUGHOUT THE INTRAERYTHROCYTIC CYCLE WITH API-AP2 TRANSCRIPT PROFILES.	219

ABBREVIATIONS

.gff	general file format
°C	degrees Centigrade
μL	microlitres
μM	microMolar
AP2	APetula2
BOC	British Oxygen Company
bps	base pairs
BSA	Bovine Serum Albumin
CDS	CoDing Sequence
DEPC	DiEthylPyroCarbonate
EDTA	EthyleneDiamine Tetraacetic Acid
ER	Early Ring
EST	Expressed Sequence Tag
EtBr	Ethidium Bromide
FIRE	Finding Informative Regulatory Elements
FL	Full Length
gDNA	genomic DNA
GEMS	Gene Enrichment motif searching
HCT	Haematocrit
HH	an intergenic distance or sequence extracted from two adjacent genes in a Head to Head (5' to 5') orientation
hpi	hours post invasion
hrs	hours
HT	an intergenic distance or sequence extracted from two adjacent genes in a Head to Tail (5' to 3') orientation
ID	gene IDentifier
IE	IntraErythrocytic
IDC	Intraerythrocytic Development Cycle
IGD	InterGenic Distance
IGR	InterGenic Region
IGS	InterGenic Sequence
iRBC	infected Red Blood Cell
l	litres
LB	Luria Broth
M	Molar
mins	minutes
mL	millilitres
mRNA	messenger RNA
NB	Northern Blot
NN	Nearest Neighbour
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline

PCR	Polymerase Chain Reaction
qPCR	quantitative PCR
RBC	Red Blood Cell
RET	Ring/Early Trophozoite
RPM	Rotations Per Minute
RT	Room Temperature
RT-PCR	Reverse Transcriptase PCR
S	Schizont
SB	Single Base
sdH ₂ O	sterile distilled water
SDS	Sodium Dodecyl Sulphate
secs	seconds
SOC	Super Optimal Broth
SSC	Standard Saline Citrate
TAE	Tris Acetate EDTA
TBE	Tris Borate EDTA
TES	Trophozoite/Early Schizont
TH	an intergenic distance or sequence extracted from two adjacent genes in a Tail to Head (3' to 5') orientation
TT	an intergenic distance or sequence extracted from two adjacent genes in a Tail to Tail (3' to 3') orientation
UTR	UnTranslated Region
v/v	volume by volume
w/v	weight by volume

ACKNOWLEDGEMENTS

I would like to extend my gratitude to my supervisor Dr. Paul Horrocks for his time, guidance and endless patience.

I would also like to thank Dr. Richard Emes, Dr. Kenneth Marx, Chai-Ho Chen and Dr. Nadia Ponts for their valuable contributions to this body of work. Dr. Richard Emes co-designed and wrote all PERL scripts that were utilized for extracting the genomic information presented in Chapter 3 and the data modelling script utilized in Chapter 5. Dr. Kenneth Marx co-designed and Chai-Ho Chen wrote all PERL scrips utilized for the homopolymeric tract positional bias work presented in Chapter 6. Dr. Nadia Ponts provided both ORF adjacent and putative transcriptional start site nucleosome positioning data for *P. falciparum*, also in Chapter 6.

Finally, last but not least, a huge thank-you to Dr. Sandra Hasenkamp and Dr. Eleanor Wong for their, support, assistance and friendship - with particular thanks to Dr. Sandra Hasenkamp with whom the functional analysis of the 5' UTR (Chapter 7) was undertaken, which she subsequently completed, when my study period came to an end.

CHAPTER 1 INTRODUCTION

1.1 MALARIA

Malaria is a disease caused by protozoan parasites of the *Plasmodium* family. These parasites are spread between human hosts by the female Anopheline mosquito vector, which requires blood meals to complete egg development. The global distribution of malaria is, therefore, driven by the range and distribution of mosquito species in which the parasite is able to complete its own sexual development. Thus, malaria is present within endemic communities as peaks and troughs of disease that directly correspond to vector abundance. Malaria, globally, affects ninety-nine countries, and whilst present within most tropical and semitropical regions of the world, the principal burden in terms of mortality falls on sub-Saharan Africa (Fig. 1-1B). Critically, the burden of malaria falls on those populations experiencing the most poverty – where inadequate housing and drainage provides for constant exposure to the mosquito vector and where detection and treatment of disease is hampered locally by access to healthcare and nationally as a result of limited health budgets (WHO, 2013).

The *Plasmodium* species; *P. falciparum*, *P. vivax*, *P. ovale* (two species) and *P. malariae* are the main etiological agents of human malaria. *P. knowlesi*, is a malarial parasite of macaque monkeys, but can also infect humans and is now commonly considered as a human malarial parasite for diagnosis and treatment purposes, (Flannery *et al.*, 2013; Singh and Daneshvar, 2013). Of these, *P. falciparum* is the most virulent and is responsible for some 90% of all malaria-related deaths (WHO, 2013). It is *P. vivax*, however, that has the widest distribution (Fig. 1-1A), and whilst previously considered benign, in more recent years it has been recognised for its importance in contribution towards *both* the morbidity and mortality (Fig. 1-1B) burdens of disease (Flannery *et al.*, 2013).

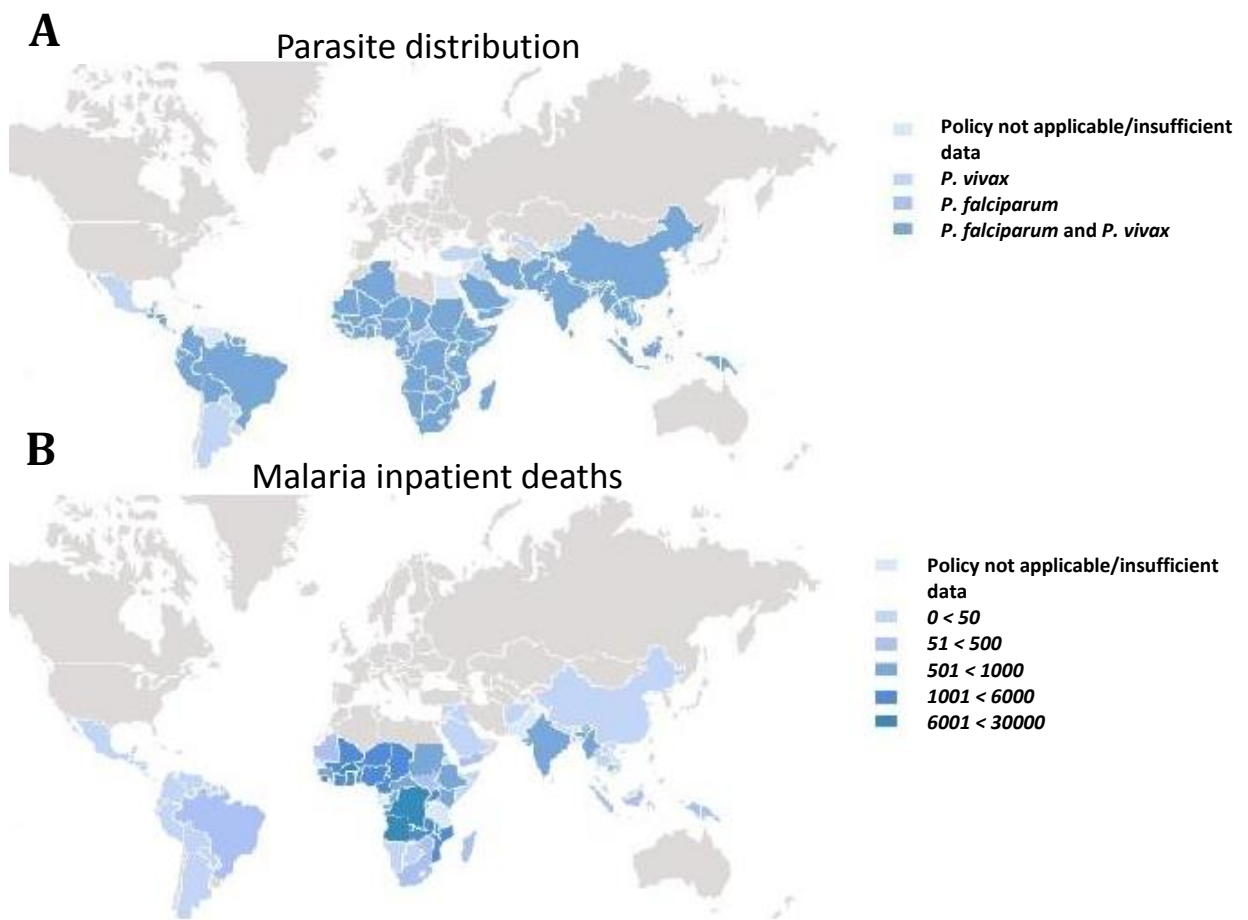


Figure 1-1 (A) Global distribution of *P. falciparum* and *P. vivax* in 2012 and (B) worldwide malaria inpatient deaths recorded in 2012.

These two figures were created from the interactive tool present on the World Health Organization website. They denote (A) the worldwide distribution of two of the most important *Plasmodium* species; *P. falciparum* and *P. vivax* respectively and (B) the distribution of malaria inpatient deaths. Of note is that the measure itself 'inpatient deaths' is likely only indicative of actual deaths in many regions of extreme poverty where those suffering from diseases such as malaria may not actually have access to any medical facility and hence the cause of their death may remain unrecorded.

In 2012 there were an estimated 207,000,000 cases of malaria worldwide resulting in at least 627,000 deaths, with some 85% of these malaria deaths in children aged under 5 years old (WHO, 2013). These figures represented a drop of almost 42% from the recent peak of estimates for malaria related mortality in 2000 and primarily result from a combination of increased awareness and funding, the wide-spread use of insecticide impregnated bed-nets,

wider availability of rapid diagnostic tests and the use of artemisinin combination therapies (Murray *et al.*, 2012; Alonso and Tanner, 2013). However, recent evidence from the World Health Organization suggests a worrying plateau in these recent declines, and it is now evident that this plateau, combined with the poor health management systems in many of the countries in which malaria is endemic, makes it likely that the optimistic Global Malarial Action Plan goals for 2015 may not be met for many of the worst affected countries (Alonso and Tanner, 2013). Additional features impact on this bleak assessment. The recent global financial crisis has resulted in a reduction in funding – where in fact additional funding is required (White *et al.*, 2014). Also, no panacea vaccine is likely to be available for malaria in the short-term, despite extensive research and development, as the leading RTS,S vaccine currently appears to lack long lasting efficacy (Agnandji *et al.*, 2011; Agnandji *et al.*, 2012). Emerging antimalarial drug resistance (such as artemisinin resistance in South East Asia) and insecticide resistance also pose a direct threat to the gains achieved since 2004 (White *et al.*, 2014), as do persistent difficulties associated with treating *P. vivax* and *P. ovale* drug-refractory hypnozoites (dormant liver-forms) - usually by the extended use of the 8-aminoquinoline Primaquine, as this can be toxic to those with glucose-6-phosphate dehydrogenase deficiency, a common ailment in endemic regions (Beutler, 1959; Flannery *et al.*, 2013). These data together suggest that the huge inroads made in combatting this disease could be lost. These challenges suggest that without further and/or new interventions, in terms of financial support and the development of replacement drugs and insecticides, that the targets established for malaria control will be hard to achieve and those for elimination even more so.

Whilst this appears daunting, some promising advances are emerging, such as the use of chemoprophylaxis for high risk groups (such as pregnant women and children under five during periods of high and intense transmission), which substantially reduces disease burden in these areas (Alonso and Tanner, 2013; White *et al.*, 2014) and attempts to develop a much

needed heat-stable version of the RTS,S vaccine (Bill and Melinda Gates Foundation in association with GlaxoSmithKline) to avoid the enormous difficulties of vaccine refrigeration in many endemic areas (London Evening Standard 29/10/2013). Vaccine development for *P. falciparum* has proved to be a significant challenge to the research community as a whole. As the intraerythrocytic (IE) cycle is primarily responsible for the disease, much research has concentrated on this life-cycle stage. Unfortunately, the monoallelic expression and antigenic variation of its major virulence factor *Plasmodium falciparum* erythrocyte membrane protein1 (PfEMP1) - coded for by the *var* genes, as well as a constantly recombining *var* gene repertoire which ensures parasite survival in the face of immune recognition and destruction, makes this an unsuitable vaccine candidate. Thus, most IE vaccine candidates have concentrated upon the antigens associated with red blood cell (RBC) invasion. There are, however, many other potential interruption points within the human host (Fig. 2). Sporozoite invasion of the liver is the first and this is considered the optimum point of intervention, although large-scale production and extraction of *P. falciparum* sporozoites from the Anopheline mosquito presents an extreme challenge. However, some opportunities for pre-erythrocytic vaccines are perhaps available through recombinant multi-subunit vaccines or genetically-attenuated parasites (Curtidor *et al.*, 2011; Vaughan *et al.*, 2010; Epstein and Richie, 2013). The development of resistance to many first-line antimalarial drugs and the recent description of the potential emergence of resistance to artemisinin in South East Asia has meant that the search for new drugs to combat malaria is now receiving urgent attention (WHO, 2013). New and novel drug candidates have been identified via massive high throughput cellular screens of compound libraries made available from large pharmaceutical companies, or via a more direct approach against crucial 'target' parasite proteins. Two such promising candidates are the spirindolones and the lead compound P218 (diaminopyridine) respectively (reviewed in Flannery *et al.*, 2013).

Antimalarial drug development, vaccine design and the search for other novel intervention therapies are predicated on our understanding of the biology, development and virulence of the malaria parasites. As a research community, there have been leaps forward in our understanding in these areas following key scientific landmarks – such as the first complete *in vitro* system to culture the IE cycle of *P. falciparum* (Trager and Jensen, 1976) and the completion of the *P. falciparum* genome project (Gardner *et al.*, 2002). These leaps, supported in between by incremental steps, provide us with an appreciation of the complex interactions that exist between the parasite, its human host and the mosquito vector. They also underpin critical targets for malaria control and elimination and disease and transmission respectively.

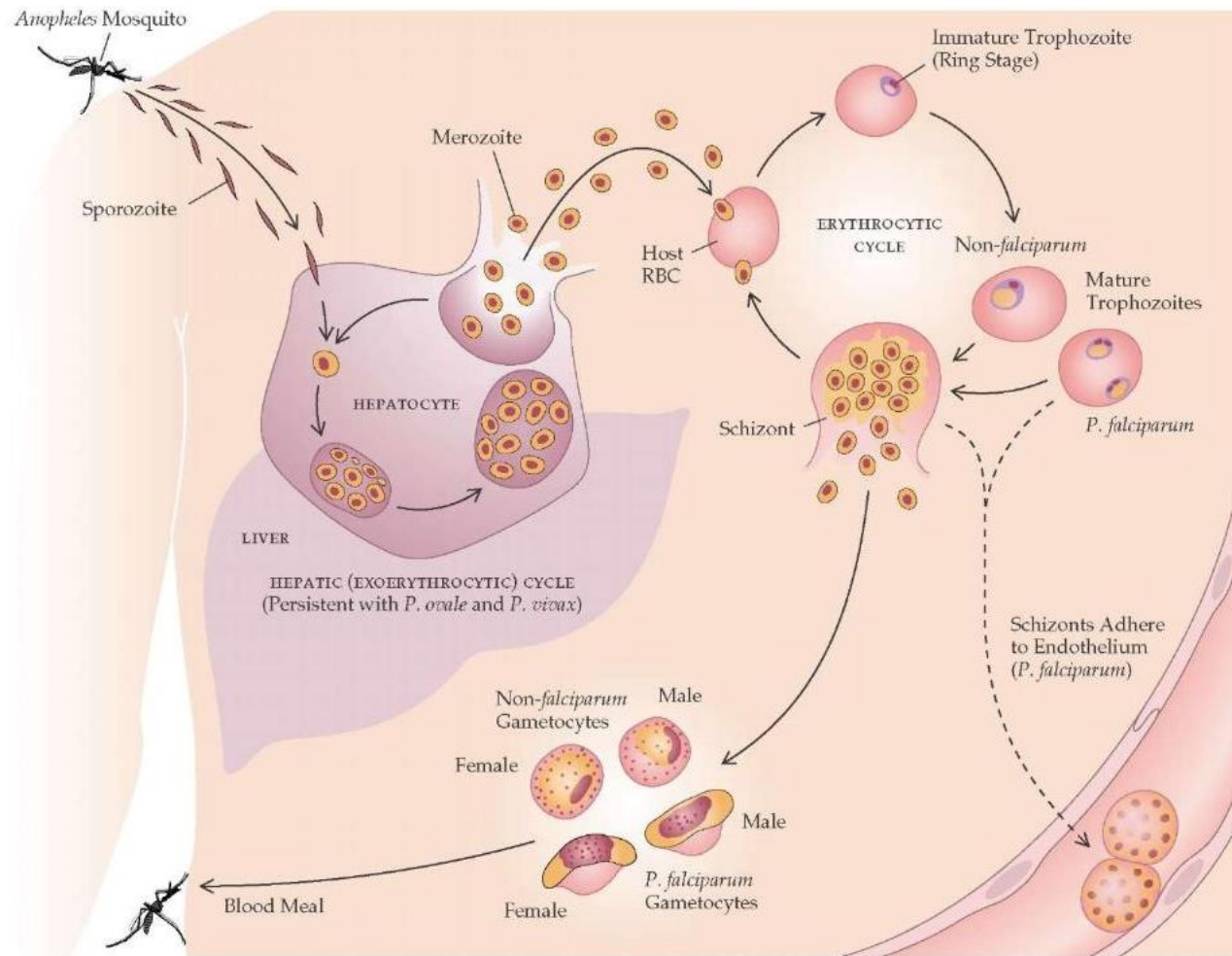
1.2 THE LIFE CYCLE OF *P. FALCIPARUM*

The life-cycle of the malaria parasite is characterized by multiple morphological stages, with each adapted to facilitate the invasion, colonisation and replication of the parasite within the diverse host cellular environments that the parasite encounters as its life cycle progresses between humans and mosquitoes. *Plasmodium spp.* essentially undergoes three replicative cycles; two asexual cycles within the human host termed exoerythrocytic schizogony and intraerythrocytic schizogony (see Fig. 1-2) and one sexual cycle which is completed within the mosquito mid-gut. The life-cycle, within the human host, starts when the sporozoite invasive forms (~15-200) are released under the skin within saliva upon mosquito blood-feeding (Vanderberg, 1977; Mota and Rodriguez, 2004). Accessing the bloodstream in the dermis, the schizonts are taken to the liver where they infect hepatocytes, likely following recognition, invasion and transition through Küppfer cells (Frevert *et al.*, 2005; Garcia *et al.*, 2006; Baer *et al.*, 2007b). Within the hepatocytes, the parasites undergo repeated rounds of replication (exoerythrocytic schizogony) over a period of 10 days (Garcia *et al.*, 2006). This period of infection is sometimes termed the latent period as it is asymptomatic. Release of the merozoite (a plasma membrane containing 100-200 merozoites) from ruptured hepatocytes

is apparently followed by transit through the heart to the lungs, for *P. yoelii* at least, where the merozoite bursts within the pulmonary capillaries and releases the merozoites into the bloodstream to initiate infection of erythrocytes (Sturm *et al.*, 2006; Tarun *et al.*, 2006; Baer *et al.*, 2007a; Vaughan *et al.*, 2008). For *P. vivax* and *P. ovale*, as mentioned previously, hypnozoites can be formed within the hepatocytes and lie 'hidden' and dormant for extended periods of time, even years, hence causing persistent infections.

Figure 1-2 The life-cycle of *P. falciparum*

(overpage) *Plasmodium* parasites in the form of sporozoites are inoculated into the human host by the bite of a female Anopheline mosquito. The sporozoites then migrate to the hepatocytes within the liver where they replicate and transform into merozoites. Merozoites then invade the red blood cells (RBC) of the infected host and traverse the intraerythrocytic asexual cycle within a species specific time-scale - ~48 hours post-infection for *P. falciparum*. As the parasite progresses through this asexual cycle within the RBC it takes on a series of morphologically distinguishable forms, primarily ring, trophozoite and schizont stages, which are readily distinguishable by light microscopy upon methanol fixation and Giemsa staining. For *P. falciparum*, infected red blood cells (iRBC) display 'knobs' of parasite protein on their surface which enables these cells to sequester within the host microvasculature resulting in host pathology. Upon maturation nascent parasites escape the ruptured iRBC and re-invade another. Gametocytes (male and female sexual forms) are produced during some cycles and these can be taken up by a female Anopheline mosquito upon blood-feeding. Sexual reproduction then takes place within the mosquito midgut. Life cycle figure taken from: <http://what-when-how.com/acp-medicine/protozoan-infections-part-1/>



Circulating merozoites rapidly (perhaps as fast as 20 seconds) locate and invade host red blood cells (RBCs) through four principal steps: recognition, reversible attachment, re-orientation and junction formation followed by secretion from specialised apical secretory organelles (micronemes, rhoptry and dense granules) which results in invagination of the membrane of the RBC and the production of a parasitophorous vacuole around the invading parasite (Cowman *et al.*, 2012). Microneme, rhoptry and dense granule activation are associated with increases in Ca^{2+} concentration, parasitophorous vacuole formation and host cell modification respectively (reviewed by (Cowman *et al.*, 2012)). Intraerythrocytic schizogony, the cyclical invasion, release and re-invasion of the host RBCs, is characterised by three principal morphological forms; the ring stage, the trophozoite and the schizont and for *P. falciparum* the IE cycle takes ~48hours to complete.

The ring stage essentially represents a growth and adaptation period, this is followed by a period of intense metabolic activity wherein haemoglobin is degraded as a source of amino acids (with its potentially toxic heme by-product stored safely within acidic food vacuoles as a polymer, β -hematin or haemozoin) and host cell cytoplasm is ingested, via the cytosome, as the parasite develops to form a mature trophozoite (Tilley *et al.*, 2011). Nuclear division (without cytokinesis), or sporogony, then occurs to produce the schizont from which ~ 16-32 nascent merozoites bud (Callan-Jones *et al.*, 2012). Essentially, the host cell is invaded and modified, whilst the parasite resides within a parasitophorous vacuole, where it develops and replicates, utilizing the resources of the host cell to enable it to propagate itself. The merozoites then egress from the infected RBC (Abkarian *et al.*, 2011) and are released into the blood stream to subsequently re-invade. This part of the life cycle is perhaps best understood given ready access to intraerythrocytic stages from a system of continuous *in vitro* culture (Trager and Jensen, 1976). Moreover, the pathology of *Plasmodium spp* in general and for *P. falciparum* in particular is almost all mediated by the intraerythrocytic stages of parasite development. The continued cycles of erythrocyte rupture can result in

profound anaemia, with associated disruption of haematopoiesis and poor nutritional status further exacerbating the anaemic state (Lalloo *et al.*, 2007; White *et al.*, 2014). Moreover, a large burden of these microaerophilic parasites generating lactic acid can lead to respiratory distress as a result of metabolic acidosis (White *et al.*, 2014). More commonly appreciated is the potential for sequestration of the mature infected erythrocytes away from circulation, and thus the risk of splenic clearance, in the peripheral microvasculature. Sequestration in the lower oxygenated regions of the circulation may facilitate parasite growth, but also offers the potential to restrict or even block normal perfusion of the organs supported by the affected microvasculature (Lalloo *et al.*, 2007; White *et al.*, 2014). Clinical presentations associated with sequestration of infected erythrocytes include cerebral and maternal-associated malaria. Rupture of erythrocytes also releases parasite-derived pyrogens that induce cytokine-mediated fever, providing the aetiological basis for the periodic fever typically associated with malaria infection (White *et al.*, 2014).

Macro- (female) or micro- (male) gametes, cell-cycle-arrested, single-nucleus sexual forms, are produced from some cycles of replication and it is the uptake of these gametocytes by the female mosquito vector that propagates onward transmission (<http://www.tulane.edu/~wiser/malaria/mal lc.PDF>). The cues for mature gamete production from the gametocytes are still not fully elucidated (Alano, 2007) but are obviously of utmost importance as it is this life-cycle form that enables malaria to be maintained within endemic communities. Gametogenesis (or sexual reproduction) is completed rapidly within the mosquito vector midgut and is thought to be triggered by factors such as; mosquito metabolites temperature reduction and an increase in carbon dioxide (Kuehn and Pradel, 2010). The parasite undergoes a series of morphological changes and development; microgamete exflagellation and formation of the zygote, with the macrogamete developing in the invasive ookinete. The maturation and traversal of the midgut epithelium by the ookinete,

resulting in the production of the replicative sexual oocyst form, culminates in the production of sporozoites which then migrate to the salivary glands (Beier, 1998; Matuschewski, 2006).

1.3 UNDERSTANDING THE CONTROL OF GENE EXPRESSION – MOVING INTO THE POST-GENOMIC ERA

*Early during the completion of this study (in 2009) I helped co-author a review providing an update outlining how advances in technology since 1999 have impacted on our understanding of the control of gene expression (attached in Appendix A). What follows in the introduction to this thesis is built around a key premise of this review – that the control of gene expression in *P. falciparum* is mediated through a combination of molecular mechanisms – but with a focus more on evidence that was just emerging in 2009 or has been published since.*

The stage-specific control of expression of the *P. falciparum* genome is critical to the parasite's ability to invade, colonise and multiply in the diverse range of host environments it encounters during its life cycle. Moreover, maintaining a chronic infection in the face of a host immune response and utilizing the host environment, often with resultant pathology, requires the ability to rapidly adapt – often by triggering progression to the next developmental stage. The study of gene expression in *P. falciparum* prior to the completion of the genome project in 2002 was essentially a series of single-gene studies, with some insights gleaned, but limited in number and impacted upon by the technology available (Horrocks *et al.*, 1998). The completion and release of the *P. falciparum* genome in 2002 (Gardner *et al.*, 2002), and other malaria parasites since (Carlton *et al.*, 2002; Hall *et al.*, 2005; Carlton *et al.*, 2008; Pain *et al.*, 2008), as well as free and ready access to data through a public repository at PlasmoDB (PlasmoDB.org) has enabled the growth and utility of the fields of functional and comparative genomics in the study of the control of gene expression.

Following the annotation of the *P. falciparum* genome, it came as somewhat of a surprise that whilst most key components of the typical basal eukaryotic transcriptional apparatus were identifiable within the *P. falciparum* genome, there was an apparent lack of recognizable general transcription factors - only a third of the normal eukaryotic arsenal (Coulson *et al.*, 2004). However, a prevalence of genes associated with mRNA stability and processing (such as CCCH-type zinc fingers) and a full complement of histone modifying and nucleosome remodelling complexes were identified (Coulson *et al.*, 2004). This observation, coupled with newly emerging transcriptomic data sets that demonstrated a temporally-linked cascade of gene expression with single peaks of mRNA abundance (Fig. 1-3) which did not appear to alter significantly with external perturbation, gave rise to the 'just-in-time' and 'hard-wired' transcription hypotheses respectively (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Gunasekera *et al.*, 2003; Llinas *et al.*, 2006; Ganesan *et al.*, 2008; Bozdech *et al.*, 2008). These data challenged the then conventional model that suggested a predominant role for promoter based *cis-trans* interactions directing the control of gene expression in this parasite - a model that was primarily driven by the limitations of genetically manipulating the parasite hence, as stated above, giving rise to studies of only a handful of genes. A functional Genomics Workshop Group met in 2006 to debate this matter - broadly dividing the evidence into transcriptional regulation and steady state mRNA levels, and post-transcriptional control. Their review of the current state of play, with key areas picked up below, moved the field towards a model where post-transcriptional mechanisms were perhaps the most critical elements of control - fuelled by the lack of apparent specific transcription factors (Deitsch *et al.*, 2007).

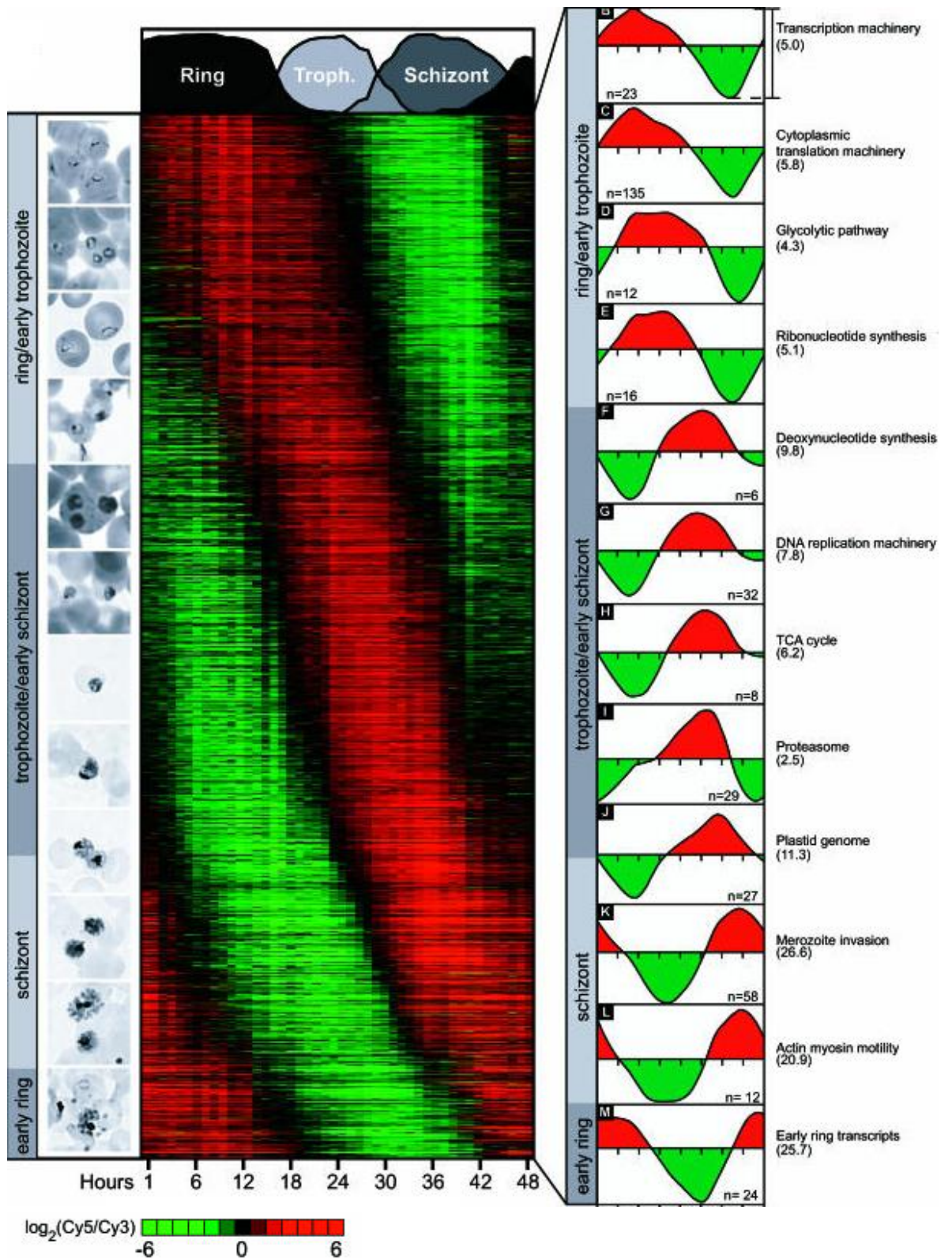


Figure 1-3 IDC Transcriptome phaseogram.

(previous page) In 2003, Bozdech *et al.* produced this elegant phaseogram depicting *P. falciparum* intraerythrocytic transcription. The intraerythrocytic developmental cycle (IDC) phaseogram is essentially a display of the temporal transcriptional profiles of 2,712 genes which have been ordered by phase of expression where representation by red is on and green is off. Bozdech *et al.*, also show the corresponding *P. falciparum* life-cycle stage on the left and the mean amplitude of expression for temporally expressed genes ordered by biochemical process or function on the right (Bozdech *et al.*, 2003). This work clearly demonstrated a predominant single maximal peak in expression for most IDC expressed genes which appeared to be ordered in a 'just-in-time' expression cascade.

Early studies exploring transcriptomic responses following exposure to small molecules identified that they appeared not to perturb transcript levels, supporting a 'hard-wired' programme of developmentally-linked transcription, although environmental perturbations had the capacity to do so (Deitsch *et al.*, 2007). Nuclear run on experiments demonstrated a sharp increase in global transcriptional activity at late trophozoite and schizont stages (with the exclusion of the *var* genes) and these data together suggested a global transcriptional initiation event, combined with post-transcriptional methods of control, could alone perhaps account for the cascade of IE gene expression demonstrated by the available microarray data (Deitsch *et al.*, 2007; Sims *et al.*, 2009; Bozdech *et al.*, 2003). Delays between mRNA and protein expression had been detected by Le Roch *et al.*, (2004) and with Shock *et al.*, (2007) demonstrating that mRNA half-lives increased from ~9.5 mins at ring-stage to ~65 mins at schizont stage, these findings together added support to this "global" hypothesis for the control of temporal gene expression (Le Roch *et al.*, 2004; Shock *et al.*, 2007). As an additional facet, translational repression had been recently identified within the gamete parasite stages in early mosquito developmental stages. This was linked to interaction with DOZI - a DEAD-box RNA helicase and a ~47bp motif located within the 3' UTR (Mair *et al.*, 2006; Braks *et al.*, 2008). These untranslated mRNA were thought to be stored within the female gametocyte (perhaps, in a manner similar to cytoplasmic messenger ribonucleoprotein complexes in higher eukaryotes) and only translated upon gametocyte activation within the mosquito vector (Mair *et al.*, 2006; Deitsch *et al.*, 2007).

This seminal meeting concluded that a series of ‘next steps’ were needed to try and steer the research community in a united manner with particular attention to chromatin organization, mRNA decay and rates of transcription, transcriptomic analysis of further life-cycle stages, comparative analysis of different species and the identification and testing of putative transcriptional regulators (Deitsch *et al.*, 2007).

1.3.1 UNDERSTANDING THE CONTROL OF GENE EXPRESSION- THE IMPACT OF TRANSCRIPTOMICS

The publication in 2003 of two separate microarray datasets that covered (i) the entire IDC in two-hour increments and (ii) seven distinct *P. falciparum* life-cycle stages demonstrated the aforementioned cascade of stage-specific and temporally-linked gene transcription with single peaks of mRNA abundance (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003). These data gave rise to the ‘just-in-time’ hypothesis of gene transcription, as co-functional transcripts appeared to accumulate together at a time immediately prior to when they were required for the parasite’s development (e.g. genes encoding DNA replication proteins were transcribed immediately before the onset of S-phase in early trophozoites). The degree of conservation of this transcriptional cascade was demonstrated from a large microarray study of three different *P. falciparum* strains (laboratory strains 3D7, HB3 and Dd2) – whereby a transcriptional shift of 12 hours and upwards was observed for only 26 mRNAs (Llinas *et al.*, 2006). Drug perturbation studies identified few transcript profile changes upon treatment with antifolates (Ganesan *et al.*, 2008) or chloroquine (Gunasekera *et al.*, 2003), supporting the ‘hard-wired’ IE program of transcription. However, changes of amplitude, increases and decreases in transcript levels, were observed with the application of artesunate to parasite culture (Natalang *et al.*, 2008) and dramatic transcriptional changes were observed 30 to 36 hours post-treatment with the choline analogue T4 (bisthiazolium compound) (Le Roch *et al.*, 2008), suggesting that *some* transcription-linked response exists to some compounds. These

studies all faced difficulties with attempting to understand whether the transcriptional changes observed were in direct response to drug perturbation or effects that resulted from induction of death and/or a quiescence in development and thus a lack of progression through the temporal cascade compared to untreated controls. In 2010 Hu *et al.*, used 20 chemical perturbations over 23 IDC time-points to establish that *P. falciparum* is indeed able to transcriptionally respond to some compounds, though to varying degrees. The most dramatic responses, consistent with developmental arrest, were from treatment with the calcium chelator EDTA, the protein kinase inhibitor staurosporine and the histone deacetylase inhibitors; trichostatin A and apicidin (Hu *et al.*, 2010). These data suggest that *P. falciparum* does retain some transcriptional plasticity and the correlation of functional gene assignment with growth-perturbation-induced transcriptional change supports the idea of co-regulation of co-functional genes . (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Hu *et al.*, 2010)

In 2008, the IDC transcriptome of *P. vivax* was published (Bozdech *et al.*, 2008). Although the *P. vivax* genome demonstrated a similar cascade of gene expression to *P. falciparum*, the mRNA abundance (2-fold change for 70% of genes) and timing of gene expression (time variations in 22% of syntenic genes) varied considerably between the two parasites (Bozdech *et al.*, 2008). This difference most likely reflects the differences between the biology of these two parasites. Of note was that peak *P. falciparum* trophozoite stage accumulation of mRNAs encoding proteins associated with cytoplasmic re-modelling and haemoglobin degradation showed dramatic changes in temporal expression profile in *P. vivax*, here being expressed primarily at schizont stages instead. This appeared to suggest some differences in the timing of re-modelling of the host cells exists between these two parasites (Bozdech *et al.*, 2008; Bozdech and Prieser, 2013).

Therefore, in summary, transcriptomic studies of *in vitro* *P. falciparum* parasite cultures demonstrate a unique, but consistent, IE expression pattern that is directly related to its life-style (and these data were supported by subsequent RNA-Seq data developed by Otto *et al.*, (2010) but appears to still retain some transcriptional plasticity in response to the external environment (Otto *et al.*, 2010).

Somewhat unsurprisingly, mRNA profiling of *P. falciparum* patient isolates produces a much more complex picture. Overall, these data together suggest that genes associated with transcriptional-flexibility tend to fall into four main categories: stress response, sexual development, energy metabolism/biosynthesis and host-parasite interactions (Bozdech and Prieser, 2013). Differential gene regulation appears to be controlled by a combination of *cis*-acting factors such as copy number variations (Mackinnon *et al.*, 2009; Gonzales *et al.*, 2008) and local chromatin organisation (Gonzales *et al.*, 2008), whereas expression level polymorphisms appear predominant and are thought more likely controlled by *trans*-acting factors (Gonzales *et al.*, 2008). Using clinical isolates for transcriptomic studies can be problematic, particularly gaining sufficient parasite mRNA for successful analyses. However, two studies using parasite isolates from Tanzania and Kenya suggest that the transcriptomes of cultured clinical isolates correlate well with 'freshly' adapted parasite samples, although the Kenyan isolate was sampled at less than 40 generations in culture (Vignali *et al.*, 2011; Mackinnon *et al.*, 2009). In the quest to understand how the transcriptomes of clinical isolates varied when compared to laboratory strains and each other, Daily *et al.*, analysed clinical isolates from Senegal and Nigeria and subsequently a much larger group from Senegal (43 patients) (Daily *et al.*, 2005; Daily *et al.*, 2007). In the 2005 study this group identified a transcriptome essentially similar to *in vitro* ring-stage cultures. Overlaid against this transcriptomic profile was an apparent increase in *stevor* and *rifin* transcription, with a somewhat smaller range of *var* gene transcription (which was probably attributable to the high polymorphism of this gene family and difficulties in microarray detection). In addition,

increased RESA-2 (putative long-chain fatty acid ligase) and PHIST (helical interspersed subtelomeric family, PF14_0752) transcription were observed (Daily *et al.*, 2005). The subsequent 2007 study enabled isolate transcriptomes to be mapped to the patient's clinical phenotype – although no significant correlations were clearly identified. However, the individual clinical isolate transcriptomes appeared to map to three physiological states; energy starvation, fermentative glycolytic growth and a much more variable group exhibiting gene expression reminiscent of environmental stress (Daily *et al.*, 2007). Another interesting observation of clinical isolates was made by Lemieux *et al.*, (2009) who identified that variation in *ex vivo* temporal culture maturation was perhaps the most significant factor in the altered transcriptomes (Lemieux *et al.*, 2009). However, a subsequent study using 58 samples from acute malaria patients from a high transmission area of Malawi, 40 of which were retinopathy-associated cerebral malaria positive, did identify distinct transcriptional profiles (cell cycle - DNA replication and invasion associated) for some of these samples which was suggestive of unique biology for severe disease (Milner *et al.*, 2012). When the transcriptional profiles from the Malawi and Senegal (taken from a low endemicity area and patients with non-severe malaria) samples were compared, two clusters, A and B, became apparent. Cluster A was correlated with low parasitaemia (containing most of the samples from Senegal) and stress whereas Cluster B was correlated with high parasitaemia and fermentative growth. This group also noted that the sequestration of *P. falciparum* late-stage parasites probably led to the predominately ring-stage transcriptome observed although some cluster A transcriptomes were more reminiscent of gametocytes and correlated well with microscopic samples (Milner *et al.*, 2012). Studies into pregnancy related malaria have identified gene-expression patterns specific to placental infection (Francis *et al.*, 2007; Tuikue Ndam *et al.*, 2008). In conjunction with apparent upregulation of *var2csa*, hypothetical proteins containing the PEXEL motif and another PHIST gene (PFI1175w) were identified, which are otherwise not detected in either peripheral blood or CSA-binding laboratory lines

and likely represent host-parasite interactions specific to this particular aetiology of disease (Francis *et al.*, 2007; Tuikue Ndam *et al.*, 2008).

Transcriptomics is also playing a role in the identification of changes in gene expression relating to drug resistance, with samples from western Cambodia (which require extended atemisinin treatment for clearance) demonstrating ring/trophozoite stage decreases in metabolic and biochemical activity and schizont-stage specific increases of protein metabolism (Mok *et al.*, 2011). These data are consistent with a 'quiescent' ring stage, resulting in lower levels of drug activation, followed by oxidative stress damage repair in schizont stages (Dondorp *et al.*, 2009; Teuscher *et al.*, 2010; Mok *et al.*, 2011). Interestingly, a transcriptomic study of glycolysis using eight different *in vivo* *P. vivax* Peruvian isolates demonstrated little change in expression for the first three enzymes of this pathway but up to 100-fold expression difference for the eight downstream enzymes (Dharia *et al.*, 2010). Therefore, dissecting changes in expression of biochemical pathways for clinical isolates may provide more useful information on disease phenotype(s).

1.3.2 UNDERSTANDING THE CONTROL OF GENE EXPRESSION- THE IMPACT OF POST-TRANSCRIPTIONAL AND POST-TRANSLATIONAL MECHANISMS

Le Roch *et al.*, (2004) documented discrepancies between protein and mRNA abundance in *P. falciparum* - predominantly in the form of a time-lag between detectable abundance of mRNA and protein. This observation was further qualified by Shock *et al.*, (2007) who, using a microarray approach on 4,000 genes, demonstrated that there was a substantial increase in average mRNA half-life during the intra-erythrocytic development cycle (IDC), increasing from ~9.5 minutes at ring stage to ~65.4 minutes at the schizont stage (Le Roch *et al.*, 2004; Shock *et al.*, 2007). No correlation was found between these data and transcript abundance or ORF length; however, correlations did exist between functional mRNA categories (Shock *et al.*, 2007). The authors surmised that regulation of the decay components themselves, and/or

additional regulatory factors including mRNA sequestration, could all contribute to this IDC related extension of mRNA half-life (Shock *et al.*, 2007). Translational repression of mRNA transcripts in female gametocytes, via the creation of a multi-subunit messenger ribonucleoprotein (mRNP) or P-body, has been identified in *P. berghei* (Mair *et al.*, 2006; Mair *et al.*, 2010). These mRNP include DOZI (RNA helicase) and CITH (Sm-like factor) and the knockout of either of the genes for these two proteins results in fertilization competence (a zygote is formed within the mosquito vector midgut), but failure to form the next developmental stage - the ookinete (Mair *et al.*, 2006; Mair *et al.*, 2010) . Therefore, 'pre-packaging' of essential transcripts for the early development stages of *P. falciparum* within the vector appears to be pre-requisite for parasite survival.

Recently, two dimensional differential gel electrophoresis (2D-DIGE); a quantitative method of protein analysis, has allowed the timing of expression and relative abundance of mRNA and protein to be compared (Foth *et al.*, 2008; Foth *et al.*, 2011) . These data paint a complex and dynamic picture and suggest that, as in other organisms, for different genes and proteins the rates of translation and protein degradation differ (Foth *et al.*, 2011). In 2008, Foth *et al.*, took schizont stage parasites and sampled them every 4 hours for a 12 hour period both via microarray and 2D-DIGE. Using 500 abundance profiles, the comparative results varied. In some cases mRNA abundance mirrored protein abundance, but in many other cases there was less correlation. These included: delay in protein abundance, a decrease in mRNA but no decrease in protein abundance, and increases in transcript but no change in protein and/or a decrease in transcript coupled with an increase in protein (Foth *et al.*, 2008). In addition, the presence of abundant protein isoforms was detected. For example, the RNA helicase eiF₄A was shown to be present in five different isoforms (Foth *et al.*, 2008). In a follow up study in 2011, using 2 hour resolution over the 48 hour IDC (again using microarray and 2D-DIGE) and a protein cross-section spanning many subcellular parasitic compartments, Foth *et al.*, demonstrated that protein profiles also exhibit a single abundance peak, a high correlation

between peak abundance for functionally related proteins and that the protein abundance tends to peak ~11 hours after its cognate mRNA abundance peak (Foth *et al.*, 2011). Critically, onset of protein expression appeared to correlate with the peak of the *rate* of mRNA transcription (Fig. 1-4). What was again starkly evident was that multiple isoforms are present for many *P. falciparum* proteins (Foth *et al.*, 2011). This group postulated that the aforementioned translational time-delay was unlikely attributable to transcriptional repression but instead more likely a combination of the dynamics of translation, degradation and isoform turnover and that these processes could be affected by a host of factors such as gene length and codon composition, poly(A) tail length, protein structure and/or susceptibility to proteolytic cleavage (Foth *et al.*, 2011). These data suggested that the small transcript changes detected during drug perturbation studies could in fact have quite significant downstream affects, albeit probably with a time delay (Foth *et al.*, 2011).

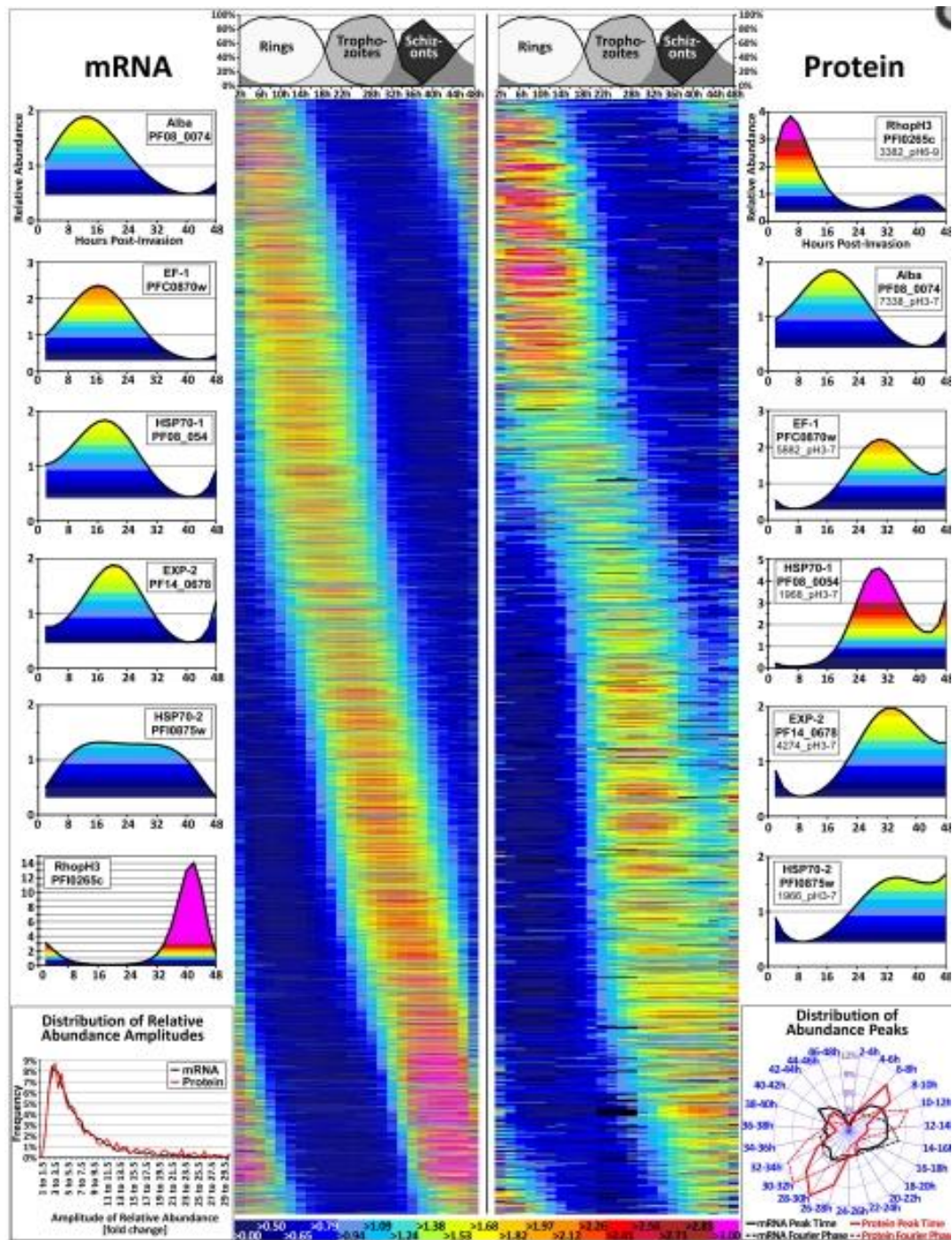


Figure 1-4 IDC Transcriptome and Proteome Phaseogram.

4,670 transcript and 1183 (single protein isoform) protein abundance profiles are shown on this phaseogram (left and right respectively) plotted, according to their Fourier phase, over the IDC. The top scale represents transition through the IDC (hours post infection) and the bottom colour scale represents abundance (transcript and protein respectively). Although in this figure transcripts and proteins are not correlated a clear cascade of protein expression - similar to that of transcript is demonstrated. As per transcript expression a predominant single peak in protein expression was observed. However, in many cases a time lag in expression was evident between transcript and protein abundance as evidenced by the six examples shown (transcript right and protein left). Foth *et al.* 2011.

1.3.3 UNDERSTANDING THE CONTROL OF GENE EXPRESSION- THE IMPACT OF PROMOTER-BASED *CIS-TRANS* INTERACTIONS

There are relatively few studies that explore the structure-function of promoter regions in *P. falciparum*, a fact attributed principally to the technical complexity of manipulating particularly AT-rich intergenic sequences. That said, the information they yield has proven invaluable in correlating the canonical structure of *P. falciparum* promoter regions with those of eukaryotes (reviewed in Horrocks *et al.*, 2009) – see also summary Table 1-1 overpage (references in figure legend). Interestingly, no consistent perturbation of the IDC temporal profile has been observed in any of these deletion studies to date, although many changes in amplitude of expression have been recorded. Subsequent work on *Pfpcna* has also been published (Wong *et al.*, 2011). These gene-by-gene studies indicate that deletion of the transcription start site (TSS) results in reporter gene abrogation, that *P. falciparum* UTR can be very long (up to an incredible 1940 and 1760 bases for *Pfmdr1* and *PfPolδ* respectively) and that enhancer and repressor sequences are likely present both upstream of the core promoter as well as within the 5' UTR (Horrocks *et al.*, 2009). Some caution, however, should be applied to these studies. Often, mis-matched promoter and terminator sequences are investigated together, with the obvious perils of interpreting sequences that do not cooperate normally. Moreover, the use of transient transfection in a significant number of studies, with the accompanying plasmid loss and absence of assembly into native chromatin, may similarly impact on the study outcome. The latter issue has been addressed by use of the *bx1* integrase system. Here genomic integration of a single copy of the reporter construct circumvents many of the issues surrounding transient and stable episomal transfection techniques (Nkrumah *et al.* 2006).

Table 1-1 Functional promoter assays in *P. falciparum*Summary of functional promoter assays in *P. falciparum*.

Gene	PlasmoDB	TSS ^a	Transfection approach	Promoter elements characterised ^b			Additional method(s) ^c	Referen
				Basal	Enhancer	Repressor		
<i>hrp3</i>	MAL13P1.480	ND	Transient episomal	☑	ND	ND	None	[53]
<i>hrp3</i>	MAL13P1.481	-626 to -750	Transient episomal	☑	☑	X	None	[190]
<i>hsp86</i>	PF07_0029	ND	Transient episomal	☑	ND	ND	None	[53]
<i>hsp86</i>	PF07_0029	-650	Transient episomal	☑	☑	X	None	[54]
<i>dhfr-ts</i>	PF0830w	ND	Transient episomal	☑	☑	X	None	[50]
<i>calmodulin</i>	PF14_0323	mult. -60	Transient episomal	☑	☑	X	None	[50]
<i>calmodulin</i>	PF14_0323	mult. -170	Transient episomal	☑	☑	X	EMSA	[46]
<i>pcna</i>	PF13_0328	-960	Transient episomal	☑	☑	X	None	[43]
<i>gbp130</i>	PF10_0159	-984	Transient episomal	☑	☑	☑	EMSA	[44]
<i>Pfs16</i>	PF0310w	-175	Transient episomal	☑	☑	X	EMSA	[42]
<i>Pfs25</i>	PF10_0303	-267	Transient episomal	☑	☑	X	EMSA	[42]
<i>PfCDP</i>	PF14_0097	-121	Transient episomal	☑	☑	X	EMSA	[191]
<i>PfPolδ</i>	PF10_0165	-1760	Transient episomal	☑	☑	☑	None	[45]
<i>Pfmdr1</i>	PFE1150w	-1940	Transient episomal	☑	ND	ND	None	[192]
<i>msp2</i>	PFB0300c	-256	Transient/stable episomal	☑	☑	☑	None	[193]
<i>var</i>	multigene	-1167	Transient episomal	☑	☑	X	EMSA	[194]
<i>rif</i>	multigene	-200 to -245	Transient episomal	☑	X	☑	EMSA	[48]
<i>Pfg27</i>	PF13_0011	-410	Stable episomal	☑	☑	X	DNase I	[58]
<i>falcipain 1</i>	PF14_0553	-466	Transient episomal	☑	☑	X	EMSA	[47]
<i>falcipain 2</i>	PF11_0165	-127	Transient episomal	☑	☑	X	EMSA	[47]
<i>falcipain 2'</i>	PF11_0161	-66 and -189	Transient episomal	☑	☑	X	EMSA	[47]
<i>falcipain 3</i>	PF11_0162	-407	Transient episomal	☑	☑	X	EMSA	[47]
<i>Pf1-cys-prx</i>	PF08_0131	-263	Transient episomal	☑	☑	☑	DNase I, ChIP	[62]
<i>Pftrx1</i>	PF14_0368	-188 to -202	Transient episomal	☑	☑	X	DNase I, ChIP	[62]
<i>rap3</i>	PFE0075c	ND	Transient episomal	☑	☑	X	EMSA	[68]

^a Transcriptional start site (TSS) in basepairs upstream of start codon. ND, not determined.^b ☑ indicates demonstration of activity, X the absence. ND, not determined.^c Additional method used to confirm *trans*-acting factor interactions. Electrophoretic mobility shift assay (EMSA), DNase I footprinting assay (DNase I) and chromatin immunoprecipitation (ChIP).

Table taken from Horrocks *et al.*, 2009. References cited in this table include; [53] Wu *et al.*, 1995, [190] Lopez-Estrano *et al.*, 2007, [54] Militello *et al.*, 2004, [50] Crabb and Cowman, 1996, [46] Polson and Blackman, 2005, [43] Horrocks and Kilby, 1996, [44] Horrocks and Lanzer, 1999, [42] Dechering *et al.*, 1999, [191] Osta *et al.*, 2002, [45] Porter, 2002, [192] Myrick *et al.*, 2003, [193] Wickham *et al.*, 2003, [194] Voss *et al.*, 2003, [48] Tham *et al.*, 2007, [58] Olivieri *et al.*, 2008, [47] Suni *et al.*, 2008, [62] Komaki-Yasuda *et al.*, 2008, [68] Young *et al.*, 2008. All references are shown in full in Appendix A Horrocks *et al.*, 2009.

Bioinformatic profiling has identified an array of putative *cis*-acting motifs using the extensive *Plasmodium spp.* genomic datasets and transcriptomic data available for *P. falciparum*, either in the form of single species studies or by taking a comparative genomics approach (Elemento *et al.*, 2007; Gunasekera *et al.*, 2007; Young *et al.*, 2008; Wu *et al.*, 2008; Jurgelenaite *et al.*, 2009). For example, the algorithm Finding Informative Regulatory Elements (FIRE) correlates over-represented motifs from sequences adjacent to ORF with transcriptomic profiles to produce a heat-map of over and under expression at different times during the IDC (Elemento *et al.*, 2007). By comparison, the algorithm Gene Enrichment Motif Searching (GEMS) pre-clusters co-functional (based on gene ontology) co-transcribed genes

prior to motif searches (Young *et al.*, 2008). Wu *et al.*, (2008), alternatively, used a comparative genomics approach utilizing *P. falciparum*, *P. knowlesi* and *P. yoelii* orthologous gene pairs which led to the identification of putative regulatory motifs within 5' and 3' ORF flanking sequence (Wu *et al.*, 2008). Therefore, a large amount of potentially useful data has been produced by using such approaches but, unfortunately, few of these motifs have been experimentally validated to date. In addition, an arbitrary length of upstream or downstream sequence has generally been selected for evaluation (usually 1000bp of flanking sequence), owing to the lack of mapped transcription start sites, which suggests that at least in some cases (e.g. *Pfmdr1* and *PfPol*) the sequences investigated may not contain the promoter region at all and that in others that this sequence length may well be truncated owing to an overlap with an upstream ORF. However, on the whole, bioinformatic approaches can be extremely useful and strongly suggest that over-represented motifs do exist within the promoter and terminator regions of *P. falciparum* genes and experimental validation of the strongest contenders should provide some interesting routes for investigation.

As stated previously, the inability to identify anywhere near a full complement of specific transcription factors (STFs) in the *P. falciparum* genome had spurred the search for other likely methods of transcriptional control in the malaria parasite (Coulson *et al.*, 2004; Deitsch *et al.*, 2007). However, the identification and subsequent characterisation of the Apicomplexan Apetela2 (Api-AP2) family of transcription factors (Balaji *et al.*, 2005; De Silva *et al.*, 2008; Lindner *et al.*, 2010) now suggests that *cis-trans* interactions over promoters do have an important role to play in transcriptional control. It has been established that out of the 26 genes (now 27) identified as encoding AP2 proteins in *P. falciparum*, 22 (now 21) of these are expressed at different time-points during intraerythrocytic schizogony - predominantly falling into four *P. falciparum* (ring, early trophozoite, early schizont and schizont) life-cycle clusters (Balaji *et al.*, 2005; Campbell *et al.*, 2010; Painter *et al.*, 2011) . The ~60 amino acid AP2 DNA binding domain in ApiAP2 is very similar to that found in plant

AP2/ethylene response factor (ERF) transcriptional regulators, however, unlike AP2/ERF, some ApiAP2 proteins contain multiple AP2 domains with few other conserved protein domains being apparent. Although relatively few AP2 containing proteins are conserved amongst all Apicomplexan species, the 27 AP2 identified in *P. falciparum* demonstrate strong conservation among other *Plasmodium spp.* (De Silva *et al.*, 2008), with >95% identity, suggesting lineage specific expansion (Campbell *et al.*, 2010). Binding specificity studies, using protein binding arrays, have identified that individual AP2 proteins bind different DNA sequence motifs, that individual AP2 domains within the same AP2 protein are capable of binding distinct motifs and that some motifs are recognized by multiple AP2 proteins (Campbell *et al.*, 2010) (Fig. 1-5). Interestingly, when these data were combined with nucleosome occupancy data (Ponts *et al.*, 2010), some 65-97% accessibility to DNA motif binding sites was predicted (Campbell *et al.*, 2010). In addition, using electrophoretic mobility shift assays this group was also able to demonstrate differing affinities for different motifs for 14 specific AP2 domains, strongly supporting a multifaceted combinatorial approach to AP2 gene regulation, such as that originally suggested by van Noort and Huynen (2006). These data are also supported by the demonstration of dimerization by the PF14_0633 AP2 protein when bound to DNA which would enable multiple motif interaction (van Noort and Huynen, 2006; Lindner *et al.*, 2010; Campbell *et al.*, 2010). These findings strongly suggest that the AP2 *trans*-acting proteins play an important role as developmental regulators during the *P. falciparum* IE cycle probably acting in concert with other regulatory mechanisms such as chromatin re-modelling machinery as the histone deacetylase inhibitor apicidin appears to alter AP2 expression profiles (Chaal *et al.*, 2010; Painter *et al.*, 2011).

PlasmoDB ID	Binding motif
PF07_0126 Tandem domain	AATTTC
PF14_0633	TGCATGCA
PF13_0267	ATTCTAGAA
PF14_0079	TGCACC
PF11_0404 Domain 1	TAAAAA
PF13_0097	TAGCTCA
PF13_0235 Domain 1	CGGGGC
PF10_0075 Domain 1	GTTCGAC
PF10_0075 Domain 2	TTGCG
PF10_0075 Domain 3	GTGCACTA
PFL1075w	TATATA
PFL1900w Tandem domain	TCTA
PFE0840c Domain 2	TGACATCA
PF14_0533	CACACAC
PFF0670w Domain 1	TAAGCC
PFF0670w Domain 2	CTCTAGAG
PFF0200c Tandem domain	GGTGCAC
MAL8P1.153	CACACA
PF11_0091	AGATAC
PF11_0442	AGCTAGCT
PFD0985w Domain 1	ACACAC
PFD0985w Domain 2	GTGTTACAC
PF13_0026	TGCACACA
PFL1085w	GTAC

Figure 1-5 List of AP2 proteins in *P. falciparum*

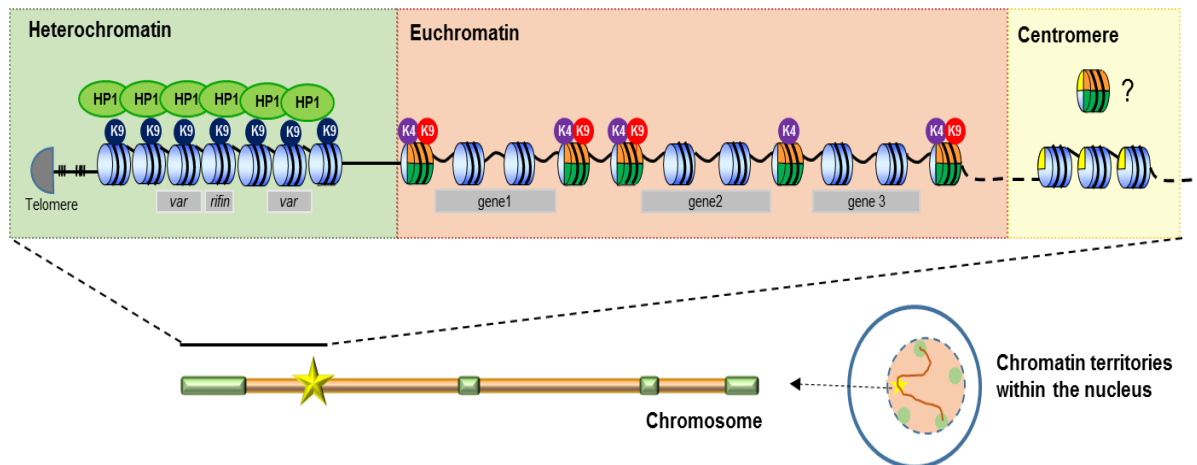
(PlasmoDB identifier provided) with their cognate predicted *cis*-motif predicted from protein binding arrays. This figure was sourced from the Malaria Parasite Metabolic Pathways website (<http://priweb.cc.huji.ac.il/malaria//maps/DNAAP2.html>) which was originally developed from data published by Painter *et al.*, 2011.

1.3.4 UNDERSTANDING THE CONTROL OF GENE EXPRESSION – THE IMPACT OF CHROMATIN ORGANIZATION AND EPIGENETIC REGULATORS

Enormous advances have been made in the field of epigenetics and chromatin structure in *P. falciparum* over the last ten years. These advances were initially driven by research into *var* gene expression and silencing, with more recent investigations expanding to complete genome wide surveys of histone modifications (Miao *et al.*, 2006; Llinas *et al.*, 2006; Trelle *et al.*, 2009; Ponts *et al.*, 2010; Westenberger *et al.*, 2009; Bartfai *et al.*, 2010; Hoeijmakers *et al.*, 2013). Epigenetics is classically defined in terms of understanding facets of gene expression that are unrelated to specific transcription factor influence or underlying DNA sequence

(Bozdech and Prieser, 2013). As such, epigenetics encompasses the packaging of DNA into chromatin which involves histones, histone variants and histone modifications (phosphorylation, methylation, acetylation, sumoylation, ubiquitination and ADP-ribosylation) to histone tail or globular domains, as well as the association of non-histone proteins and non-coding RNAs. Chromatin normally resides in either a compacted state (heterochromatin) or a 'looser' transcriptionally competent state (euchromatin) (Fig. 1-6) but is a dynamic structure and is readily modified by both covalent and non-covalent means (Westenberger *et al.*, 2009; Bartfai *et al.*, 2013). The *P. falciparum* nucleosome contains the canonical histone octamer; two H2A/H2B dimers and one H3/H4 tetramer with ~ 150bp of DNA in two superhelical turns wrapped around it (Bartfai *et al.*, 2013). The *P. falciparum* genome also encodes H2A.Z, H2B.Z (Apicomplexan-specific), H3.3 and CenH3 variant histones but does not appear to contain the linker histone H1 which probably explains why the intense metaphase chromosomal compaction is not apparent in this parasite (Bartfai *et al.*, 2013; Ponts *et al.*, 2010; Hoeijmakers *et al.*, 2012a; Hoeijmakers *et al.*, 2013). The *P. falciparum* genome also contains histone methyltransferase, demethylase and acetyl transferase homologues and a heterochromatin protein 1 (HP1) homologue (responsible for silent heterochromatin propagation). Although previous work in *P. falciparum* had been unable to identify methylated nucleosides (Choi *et al.*, 2006) a recent publication by Ponts *et al.*, (2013) has identified a *P. falciparum* methyltransferase (PfDNMT) responsible for asymmetric cytosine methylation (Me⁵C) - more akin to that found in undifferentiated mammalian and plant cells (Ponts *et al.*, 2013) than classic eukaryotic CpG methylation. Therefore, most of the required apparatus for epigenetic control, or contribution to control, of gene transcription in *P. falciparum* are present.

Figure 1-6 Chromatin domains of the *P. falciparum* genome



The *P. falciparum* epigenome can be broadly categorized into three chromatin domains: heterochromatin (depicted green), euchromatin (light brown) and centromeres (yellow). These domains occupy different parts of the nuclear space as well as the chromosomes (lower panels) and carry distinct epigenetic marking (upper panel). Genes are depicted as grey boxes, histone variants are drawn in different color (canonical histones are blue, H2A.Z is orange, H2B.Z is green while CenH3 is yellow). Modifications of histone H3 are indicated by small balls on top of the nucleosomes (H3K9me3 in dark blue, H3K9ac is red while H3K4me3 is purple). This figure was kindly provided by Richard Bártfai from an online Malaria encyclopedia article entitled “Chromatin structure and functions” authored by Bártfai, Cui, Horrocks and Miao.

The enigma surrounding the clonal expression of a single *var* gene variant and silencing of all other *var* variants has prompted considerable research into control of this phenomenon. It has been demonstrated that *var* gene silencing is associated with H3K9me3 promoter enrichment (Chookajorn *et al.*, 2007; Lopez-Rubio *et al.*, 2007) the presence of the heterochromatin protein 1 (PfHP1) which binds specifically to H3K9me3 (Flueck *et al.*, 2009; Perez-Toledo *et al.*, 2009) and the silent information regulators *PfSirA* and *PfSirB* (NAD⁺-dependant histone deacetylases or sirtuins) which repress *var* gene transcription by hypoacetylating heterochromatin (Freitas-Junior *et al.*, 2005; Duraisingh *et al.*, 2005). It has been established that the *PfAlba3* protein, which can be deacetylated by *PfSir2A*, is also subtelomerically co-located (Goyal *et al.*, 2012). Non-coding RNAs have also been implicated in the maintenance or formation of heterochromatin (Broadbent *et al.*, 2011) and, at schizont stage, *PfSIP2* was demonstrated to specifically bind *upsB*-type upstream SPE2 arrays from the vicinity of which telomerically transcribed long non-coding RNAs (LncRNA-TARE) were

produced (Flueck *et al.*, 2010; Sierra-Miranda *et al.*, 2012) suggesting that their transcriptional activation could be linked to *PfSIP2* binding (Hoeijmakers *et al.*, 2012b). Another important facet of *var* gene silencing is nuclear sub-location which was first demonstrated for *P. falciparum* in 2005 using DNA fluorescence in situ hybridisation (FISH) (Ralph *et al.*, 2005; Voss *et al.*, 2007). The active movement of a transcriptionally silent *var2csa* gene from a subtelomeric cluster on the nuclear periphery to a different nuclear peripheral site and subsequent transcriptional activity was then demonstrated (via RNA-FISH) by Lopez-Rubio *et al.*, in 2009. Essentially, heterochromatic peripheral clusters (usually 4 or 5) are apparent in the nucleus for the *Plasmodium* chromosome ends (Freitas-Junior *et al.*, 2005; Flueck *et al.*, 2009; Hoeijmakers *et al.*, 2012b) with the active *var* gene residing in a separate transcriptionally permissive nuclear zone (reviewed by Deitsch and Dzikowski, 2013). Interestingly, not only *var* genes are subject to specific nuclear compartment localisation: H3K9 methylation is carried out by histone lysine 9 methyltransferase (*PfKMT1*) and both *PfKMT1* and H3K9me3 enrichment were co-located to the same perinuclear space (Cui *et al.*, 2008; Lopez-Rubio *et al.*, 2009). *PfHP1* and *PfSIR2* were also shown to be co-located at the nuclear periphery within the electron-dense heterochromatic region (Freitas-Junior *et al.*, 2005; Flueck *et al.*, 2009). These data all correspond well with the anticipated location of the transcriptionally silent *var* genes. The active *var* gene by contrast, as stated above, resides in a separate transcriptionally permissive nuclear zone and is associated with the incorporation of the variant histone H2A.Z, at least at ring stage parasites (Petter *et al.*, 2011), and is associated with the histone marks H2K4me2, H2K4me3, H3K9ac (Lopez-Rubio *et al.*, 2007; Volz *et al.*, 2012) and H4 acetylation (Freitas-Junior *et al.*, 2005). *PfMYST* (an H4 acetyl transferase) has also been associated with active *var* promoters (Miao *et al.*, 2010) as has *PfSET10* (a H3 lysine 4 methyltransferase) during later IE development (Volz *et al.*, 2010), although it is thought that the role of *PfSET10* may be more linked to *var* gene 'memory' (Volz *et al.*, 2010; Hoeijmakers *et al.*, 2012b). Therefore, research into monoallelic

var gene expression has contributed substantially to our understanding of some of the epigenetic marks associated with *P. falciparum* gene transcriptional activation and silencing.

Whole genome studies using DNA amplification followed by mass parallel sequencing of immunoprecipitated chromatin (ChIP-Seq) and chromatin immunoprecipitation linked with microarray detection (ChIP-on-chip) have also produced interesting results, enabling changes in epigenetic markers to be tracked during the IE cycle, including positional changes, variable abundance and specific genomic positioning (Bozdech and Prieser, 2013). These data combined paint a picture of a divided chromosome, with a mainly euchromatic centre and heterochromatic centromeres and telomeric/subtelomeric regions (Fig. 1-6). These different domains are identifiable by a series of epigenetic marks that dynamically alter throughout the *P. falciparum* IE cycle and operate within a complex nuclear architecture with discrete domains for transcriptional activity and repression. Two studies were published in 2009 and 2010 on the dynamics of nucleosome positioning during the IDC (Westenberger *et al.*, 2009; Ponts *et al.*, 2010). Ponts *et al.*, (2010), used a combination of formaldehyde-assisted isolation of regulatory elements (FAIRE), to detect protein-free DNA, and micrococcal nuclease digestion of mononucleosomes (MAINE), to detect histone bound DNA, along with massively parallel sequencing during 7 *P. falciparum* life-cycle time-points (6 hour increments). The data from these two complimentary methodologies suggested that during most of the IE cycle, the *P. falciparum* genome contains more protein-free than histone-bound DNA. A nucleosomal preference for coding regions was also detected along with an apparent massive nucleosomal depletion at early trophozoite stage, with the exception of telomeric and subtelomeric regions, followed by a gradual repositioning between 24 and 36 hours post-invasion (hpi). Promoter regions appeared to be nucleosome free; 3 types of promoter region were identified each with a differing nucleosome free upstream region (Type 1 -400 to 0, Type 2 -650 to -200 and Type 3 -1000 to -600). Analysis of chromatin status over time (by K-means clustering) provided evidence of an open cluster of genes at ring stage (I), whereas

two clusters were open at late trophozoite stage (II, III), with all three of these clusters closing again at schizont stage. The fourth cluster appeared to be partially open at ring stage, be packed again at the trophozoite stage then partially re-opened at 36hpi. These data suggest a model of a set of ring-stage transcribed genes (I), and the gradual opening and re-packing of the majority of the rest of the genome between trophozoite and schizont life-cycle stages, respectively (Ponts *et al.*, 2010). Interestingly, the nucleosomal depletion observed within intergenic regions was interpreted by the authors as a possible consequence of the extreme AT-richness of *P. falciparum* intergenic regions and the much lower nucleosome affinity previously observed for poly dA.dT tracts owing to their structural and mechanical attributes (Segal and Widom, 2009b; Ponts *et al.*, 2010). The second report (Westenberger *et al.*, 2009) which used ChIP-chip (immunoprecipitated with the anti-H4 antibody) and Affymetrix microarray again suggested that the intergenic regions of *P. falciparum* were relatively nucleosome free, that there was no correlation between nucleosome occupancy and mRNA levels, apart from the *var* genes which exhibited a negative correlation, also that nucleosome positioning during the IDC appeared to be a dynamic process (Westenberger *et al.*, 2009). However, the publication of a subsequent report by Bartfai *et al.*, (2010), strongly suggests that these euchromatic intergenic regions are in fact populated by nucleosomes but, that they contain the *PfH2A.Z* histone variant and that the localization of the *PfH2A.Z* variant correlates well with the H3K4me3 and H3K9ac euchromatic epigenetic marks. Furthermore, this group also described a double-variant nucleosome type comprised of *PfH2A.Z* and the Apicomplexan-specific *PfH2B.Z* which demarked euchromatic regions, expressing a particular deposition preference for AT-rich regions of DNA - which are prevalent within intergenic regions (Bartfai *et al.*, 2010; Hoeijmakers *et al.*, 2013). They speculate that perhaps, as H3K4me3 and H3K9ac epigenetic marks demonstrate a more dynamic IDC profile, that H2A.Z and/or H2B.Z may act as a recognition point(s) for enzymatic addition or removal of these epigenetic marks, that nucleosomal properties could be influenced by extensive acetylation of

PfH2B.Z (Miao *et al.*, 2006; Trelle *et al.*, 2009), or that as these double-variant nucleosomes were likely less stable, they may allow access to the DNA by transcriptional machinery or chromatin remodellers (Hoeijmakers *et al.*, 2013). One particularly interesting outcome of this research was the positive correlation between genomic AT-content and *PfH2A.Z/PfH2B.Z* nucleosomes suggesting an AT-rich DNA preference for this double-variant nucleosomal type (Hoeijmakers *et al.*, 2013). Hoeijmakers *et al.*, (2012), in their review, suggest that the disparity between their data and previous published data on the nucleosomal landscape of intergenic regions (Westenberger *et al.*, 2009; Ponts *et al.*, 2010), may well be a product of the extreme AT-content of intergenic regions (up to 90% AT) and the choice of the amplification methodology employed or, perhaps the intrinsic reduced stability of this histone double variant type which could have resulted in sample handling displacement (Hoeijmakers *et al.*, 2012b).

The use of a combinatorial histone code is also suggested in a recent report by Gupta *et al.*, (2013), whereby utilizing a series of histone acetylation and methylation 'marks' the dynamics of euchromatic regions are probed for links to particular chromosomal domains i.e. IGR or ORF, genetic loci over the *P. falciparum* life-cycle and positive or negative correlations with transcription respectively. Thirteen histone 3 or 4 marks were used in total; 7 lysine acetylation marks, 1 tetra acetylated lysine (H4ac4 – acetylated at 4,8,12 and 16) and 5 lysine or arginine methylation marks. H3K4me3 was confirmed to demark non-coding regions and appears uncoupled from transcription whereas H4K9ac appears to exhibit positional bias for promoter regions - particularly in active genes. H4K16ac, H4ac4, H3K56ac, H3K9ac, and H3K4me3 also appear to be associated with transcription, but these marks are positively correlated with the 5' ends of ORFs. Interestingly, H4K5ac is negatively correlated and has been proposed to be a transcriptional repressor (Gupta *et al.*, 2013). In addition, variable occupancy, to some degree, for at least some of their genetic loci, was demonstrated for all of these histone marks during the *P. falciparum* life-cycle. For example H4K8ac and H3K4me3 at

>50% of their loci demonstrated significant changes in signal (Gupta *et al.*, 2013). Of particular interest was the dynamic occupancy of genetic loci for different histone marks at different *P. falciparum* life-cycle stages whereby at 0-16hpi H4K20me1, H4k20me3 and H3K14ac were maximally occupied at > 50% of their genetic loci whereas maximal occupancy at 16-32hpi was by H3K4me3 and H4K5ac (40% of their genetic loci) and at late schizont stages H4K8ac, H3K9ac (>40% of their genetic loci) (Gupta *et al.*, 2013). These data suggest that the marking of histones is also a dynamic process and associated with life-cycle stage and *de novo* transcription itself. The fact that so few histone marks were associated specifically with IGR and promoter regions, but that many were associated with the 5' end of ORF and correlated with gene transcription, led these authors to surmise that perhaps *P. falciparum* chromatin structure, as well as its transcription factors (ApiAP2), may be more reminiscent of plants than eukaryotes (Gupta *et al.*, 2013).

Therefore, most of the central part of each *P. falciparum* chromosome is euchromatic in nature (with the exception of the centromeres which have been shown, during mitosis and cytokinesis, to condense within their own heterchromatic nuclear 'pore' (Hoeijmakers *et al.*, 2012b). However, coding sequence and intergenic sequence exhibit different epigenetic marks (Hoeijmakers *et al.*, 2012b; Gupta *et al.*, 2013), with many of these marks present within the extreme 5' region of the ORF rather than within the IGR itself (Gupta *et al.*, 2013). Interestingly, the H3K20me2 mark, normally a silencing mark, has also been associated with euchromatic domains in *P. falciparum* (Lopez-Rubio *et al.*, 2009) as have the two histone acetyltransferases *PfGCN5* (H3K9 and H3K14) and *PfMYST* (H4 acetylation) and a histone deacetyltransferase inhibitor suberoylanilide hydroxamic acid (SAHA) all of which seem essential for parasite viability (Cui and Miao, 2010; Miao *et al.*, 2010). The relative depletion of the euchromatic H3K4me3 and H3K9ac marks, apparent at the ring stage of development, have also given rise to theories of 'BLACK' chromatin (devoid of classical activation or deactivation marks) - as identified in *Drosophila* (Filion *et al.*, 2010). More detail on

chromatin structure and function in *P. falciparum* is provided in the recent review of chromatin structure and function which tabulates core histones and variants, histone post-translational modifications (PTM) and proteins containing PTM domains identified in this parasite to date (Bartfai *et al.*, 2013). Nuclear architecture, which appears to play such an important role in *var* gene expression, is also under extensive research using 'state-of-the-art electron microscopy' (Eshar *et al.*, 2011; Weiner *et al.*, 2011). The nucleus of ring stage parasites appear to show little distinction between heterochromatic and euchromatic domains however, trophozoite and schizont stages show distinct nuclear chromatic regions (Hoeijmakers *et al.*, 2012b) suggesting that the changing 3-D nuclear architecture corresponds to parasite life-cycle transcriptional demands (Hoeijmakers *et al.*, 2012b; Weiner *et al.*, 2011; Sims *et al.*, 2009) and thus plays an important role in controlling *P. falciparum* transcription.

In summary, the advances that have been made in this field in particular, over a relatively short research period, have been immense; however, a clearer understanding of the enzymatic 'orchestrators' and 'pathways' of the dynamic epigenome are fundamental (Hoeijmakers *et al.*, 2012b).

1.4 CONTROL OF GENE EXPRESSION IN *P. FALCIPARUM* IS MEDIATED AT ALL LEVELS OF TRANSCRIPTION AND TRANSLATION – AND RELIES ON SIGNALS IN THE INTERGENIC REGIONS

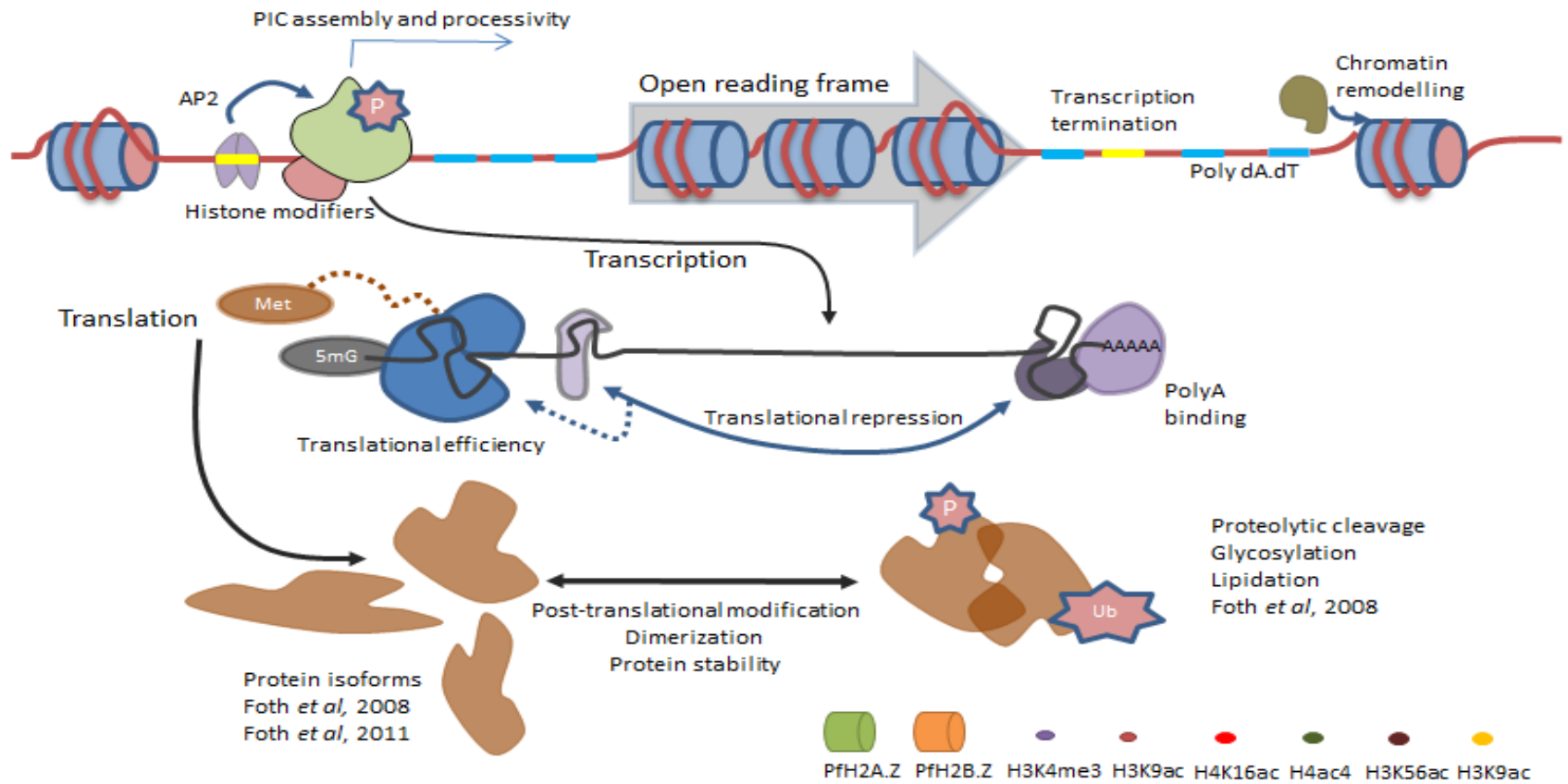
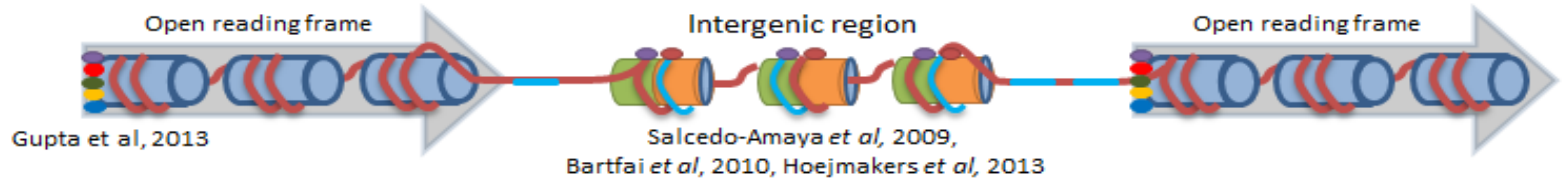
Investigations into the molecular mechanisms that govern the stage-specific and monoallelic control of gene expression have, since 2002, extensively expanded our knowledge base to uncover a complex dynamic picture of the multi-faceted approach *P. falciparum* employs to carefully control its complex life-cycle. In summarising what is known to date we see (i) *P. falciparum* chromosomes are divided into particular chromatin domains with

heterochromatic centromeres and subtelomeric regions which appear to localize to specific clusters within the nuclear periphery whilst the rest of the chromosome remains in a relatively uncondensed euchromatic state. Using Fig. 1-7 to illustrate this, euchromatic regions are characterized by the presence of *PfH2A.Z/PfH2B.Z* nucleosomes within intergenic regions, H3K9ac epigenetic marks at promoters and H4K16ac, H4ac4, H3K56ac, H3K9ac, and H3K4me3 epigenetic marks within the 5' portion of the ORF. (ii) Transcription initiation within this environment appears reminiscent of a classical eukaryotic bipartite promoter, utilizing monocistronic transcription by RNA Pol II, which may reside in a 'poised' state within the intergenic regions (Gopalakrishnan *et al.*, 2009; Levine, 2011) probably controlled via the interaction of *trans*-acting factors such as the ApiAP2 proteins, either via direct interaction with promoter based *cis*-acting sequences, via accessibility to nucleosome free TSSs, or perhaps more likely through thermodynamic equilibrium of all factors involved (Segal and Widom, 2009a; Segal and Widom, 2009c). (iii) However, the location of many epigenetic marks within the the 5' start of ORFs, rather than within the IGR itself, combined with life-cycle histone modification dynamics data strongly suggests that chromatin remodelling may also play an equally important role. (iv) Production of the transcript appears to be at one predominant time-point during the IE life-cycle and the half-life of most transcripts increases throughout the life-cycle. (v) Production of protein also appears to be at one predominant time-point during the IE life-cycle, but temporal transcript abundance does not generally correlate with protein abundance - most commonly exhibiting a time lag between maximal transcript expression and maximal protein expression. The balance between mRNA decay and protein production and decay does not appear to be a 'hard-wired' process as it appears different for different genes or sets of functionally related genes and this may be partially attributable to multiple protein isoforms from the same ORF - the transcripts from which may be more or less stable and/or the longevity of the protein may be influenced by a myriad of post-translational modifications - both of which likely play

important roles in protein dynamics. Unfortunately, little information is currently available on methods of mRNA stability or decay or the process of translation and its control in *P. falciparum*.

Figure 1-7 Overview of regulation of gene expression in *P. falciparum*

(overpage) This figure provides a schematic from current literature summarising the multiple layers of molecular control that direct gene expression in *P. falciparum* to date (note these data are applicable only to chromosome internal regions not centromeric or subtelomeric regions). Also, this figure is by no means exhaustive. But, hopefully provides a visual overview of the transcriptional and translational processes and many of the factors that are thought to influence control of transcription, translation and protein stability in this parasite.



Molecular mechanisms such as epigenetic mechanisms that provide access to genetic factors encoded within the genomic template, binding of basal and regulatory components of the RNA metabolic machinery necessary for mRNA production, facets of the mRNA UTR involved in maturation, stability and ultimately translation - all operate predominantly over intergenic regions (IGR) - a compartment of the *P. falciparum* genome that is relatively poorly characterized. Given that ~53% of the genome encodes protein it is likely that some 44-47% of the genome lies within the IGR - this reflects the relatively compact organisation of the genes in this organism (1 gene every *c.*4600bp). Evidence for the importance of these regions in the control of RNA metabolism are apparent from studies of evolution, selective constraint and selective forces acting upon the non-coding IGR of related *Plasmodium spp.* (Neafsey *et al.*, 2005; Essien *et al.*, 2008; Nygaard *et al.*, 2010). Selective constraint of upstream regions flanking many genes, particularly those involved in sexual development, cell invasion and chromatin assembly (i.e. CCCH zinc finger, ion transport and transcription factors) is high across *Plasmodium spp.* - in many cases much higher than within their cognate coding regions (Essien *et al.*, 2008). Selective constraint is also evidenced by the presence of specific conserved elements, not only within coding regions, but also within intergenic and intronic regions across *Plasmodium spp.* (Nygaard *et al.*, 2010) suggesting the presence of an evolutionary footprint of conserved elements necessary for RNA metabolic processes.

Critical to characterizing the molecular mechanisms that govern temporal transcription is an understanding of the key transcriptional landmarks - transcription start and stop sites. Defining these positions provides a context to appreciate the location of regulatory regions adjacent to and within the UTR. Transcription start and termination sites are poorly defined in *P. falciparum* - to a large extent due to the extreme AT bias within intergenic sequences that can typically exceed 90%. In addition, the presence of large homopolymeric poly dA.dT tracts in these regions makes them difficult to manipulate in functional analyses and also makes bioinformatics approaches challenging. EST databases exist for *P. falciparum*;

however, EST data generated using either oligo dT or 5' RNA ligase mediated rapid amplification of cDNA ends (RLM-RACE) each face limitations. Oligo dT rarely extends into the 5' UTR and may mis-prime from homopolymeric poly dT tracts within the intergenic regions. Available 5' RLM-RACE data (available for 1465 ORF) suggests the 5' UTR is between 150-450bp upstream of the ORF – often with multiple start sites (Watanabe *et al.*, 2001). These data are at apparent odds with available northern blot data which suggests that UTR are much larger and emanate from only one or two sites - although northern blot data is also limited (see also chapter 4 Introduction). In 2007, Brick *et al.*, used a bioinformatic approach to predict core promoters based on the physiochemical properties of DNA. This algorithm was trained using the RLM-RACE data and did, apparently, extend the distance of core promoters further upstream of the ORF (Brick *et al.*, 2008). Finally, some functional studies have physically or functionally (promoter deletion studies) mapped the TSS (reviewed in Horrocks *et al.*, 2009) and these experimental studies suggest predominantly a single TSS some 400-1900bp upstream of the ORF. Therefore, considerable disparity is apparent between all of these datasets which is difficult to resolve without further data.

1.5 AIMS OF MY RESEARCH

To date, there has been no systematic exploration of the organisation of intergenic regions or how this space correlates with what is known about transcription over these intergenic regions (RNASeq and microarray data is essentially only available for ORF). Therefore, for my PhD studies, I set out to address the following questions;

1. How are intergenic regions organised over the entirety of the *P. falciparum* genome and is this organisation distinct in different chromosomal compartments?
2. Given that a significant amount of the mRNA transcript is likely untranslated, how are these transcripts organised over the available intergenic space?

3. Poly dA.dT tracts would be expected to be overrepresented in the AT rich genome of *P. falciparum*, but is this actually the case, and how are these poly dA.dT tracts organised with respect to landmarks for expression and nucleosome occupancy data?
4. Can the genome data for other *Plasmodium spp.* in particular, and other Apicomplexan parasites in general, be used to support answers to the questions posed above?
5. Given the potential for long 5' untranslated regions within the mRNA transcript, what effect do deletions of these long 5' UTRs have on the absolute and temporal control of expression of a prototypical *P. falciparum* house-keeping gene.

Work described towards these aims or as a review of the literature has been published in the following articles:

Appendix A

Horrocks P, Wong E, Russell K, Emes RD. Control of gene expression in *Plasmodium falciparum* - ten years on. *Mol Biochem Parasitol.* 2009, 164(1):9-25. doi: 10.1016/j.molbiopara.2008.11.010.

Appendix B

Hasenkamp S, Russell K, Ullah I, Horrocks P. Functional analysis of the 5' untranslated region of the phosphoglutamase 2 transcript in *Plasmodium falciparum*. *Acta Trop.* 2013, 127(1):69-74. doi: 10.1016/j.actatropica.2013.03.007.

Appendix C

Russell K, Hasenkamp S, Emes R, Horrocks P. Analysis of the spatial and temporal arrangement of transcripts over intergenic regions in the human malarial parasite *Plasmodium falciparum*. *BMC Genomics.* 2013, 14:267. doi: 10.1186/1471-2164-14-267.

CHAPTER 2 MATERIALS AND METHODS

2.1 LABORATORY STOCK REAGENTS

Note: chemicals were sourced from either Sigma (UK) or VWR (UK) unless stated otherwise.

Ethylene Diamine Tetraacetic Acid (EDTA) 0.5M, pH8.0

93.05g of disodium EDTA.2H₂O was added to 400ml sterile distilled water (sdH₂O) and placed on a magnetic stirrer. 10g of NaOH pellets were added, the solution adjusted to pH8.0 (using 10N NaOH) and then made up to 500ml with sdH₂O. The solution was then autoclaved to sterilise.

10X Tris-acetate-EDTA (TAE) buffer

48.4g Tris (hydroxymethyl) methylamine, 11.42ml glacial acetic acid and 20ml 0.5M EDTA (pH8.0) were added to sdH₂O to total volume of 1 litre. The solution was autoclaved to sterilise. Used finally at a X1 concentration.

5X Tris-borate-EDTA (TBE) buffer

54g Tris (hydroxymethyl) methylamine, 27.5g boric acid and 20ml 0.5M EDTA (pH8.0) were added to sdH₂O to total volume of 1 litre. The solution was autoclaved to sterilise. Used finally at a X1 concentration.

20X SSC (3.0M NaCl, 0.3M sodium citrate)

175.3g NaCl and 88.2g sodium citrate were added to 800ml sdH₂O, the solution adjusted to pH7.0 (using 14N HCl) and then made up to 1 litre with sdH₂O. The solution was then autoclaved to sterilise and diluted as required.

Gel Loading Buffer X6

0.25% (w/v) bromophenol blue solution, 0.25% (w/v) xylene cyanolFF and 30% (v/v) glycerol in sdH₂O. This was stored at 4°C and used at 1 x concentration.

Church Hybridisation Buffer (0.5M NaP pH7.2, 7% SDS, 2% Dextran sulphate, 1mM EDTA)

20.7g NaH₂PO₄, 4.5gNaOH and 0.6ml 0.5M EDTA were added to 270ml sdH₂O which was mixed on a heated magnetic stirrer. The solution was then adjusted to pH7.2 (using 10N NaOH) and 21g SDS and 6g dextran sulphate were added and enough sdH₂O to make the solution up to 300ml. When completely dissolved the pH was then checked and re-adjusted if necessary.

LB Medium

10g tryptone, 5g yeast extract and 10g NaCl were dissolved in 1 litre of sdH₂O. The pH was adjusted to pH7.0 using 5N NaOH then the medium was autoclaved and stored at 4°C.

LB Agarose

15g bacteriological agar was added to 1 litre of LB medium prior to autoclave and stored at room temperature until required. To make LB agar plates, the LB agar was melted in a microwave and allowed to cool to approximately 50°C. Ampicillin (at a final concentration of 50µg/ml) was added and the plates poured into petri plates and allowed to set. Once set, the plates were inverted and stored at 4°C. Plates were warmed at 37°C to remove condensation prior to use.

SOC Medium

20g tryptone, 5g yeast extract and 0.5g NaCl were added to 950ml sdH₂O, the solution was mixed on a magnetic stirrer. 10ml of 250mM KCl was added and the pH adjusted to pH7.0

with 5N NaOH. The solution was then made up to 1 litre with sdH₂O and autoclaved to sterilise. Prior to use 5ml of sterile 2M MgCl₂ was added.

WR99210

25mM stock aliquots were stored at -20°C in DMSO (gift from Jacobus Pharmaceuticals, Princeton, NJ). To 1ml of *P. falciparum* culture medium 1µl of 25mM WR99210 was added to give a 25µM concentration (stored at 4°C for 1 week). 100µl of the 25µM stock was added to 500ml *P. falciparum* culture medium to give a final concentration of 5nM.

Blasticidin S hydrochloride

To 5ml of sdH₂O 0.05g Blasticidin S hydrochloride (Invitrogen, UK) was added to give a 10mg/ml stock. This was stored in 1.25ml aliquots at -20°C. 125µl was added to 500ml *P. falciparum* culture medium to give a final concentration of 2.5µg/ml.

G418 (Neomycin)

To 2.5ml of sdH₂O 0.5g G418 was added to make a 200mg/ml stock. This was stored at 4°C. To 500ml *P. falciparum* culture medium 312.5µl of the 200mg/ml stock was added to give a final concentration of 125µg/ml.

Cytomix

6mL 2M KCl, 7.5µl 2M CaCl₂, 1ml of K₂HPO₄/KH₂PO₄ pH 7.6 (for 10ml of 1M phosphate buffer pH 7.6 mix 8.66ml of 1M K₂HPO₄ and 1.34ml of 1M KH₂PO₄), 10ml of 250mM Hepes/20mM EGTA pH7.6 (pH to 7.6 using KOH) and 0.5ml of 1M MgCl₂ were mixed together in a total volume of 100ml with sdH₂O. This was stored at 4°C for up to 2 months.

2.2 CULTURE OF *PLASMODIUM FALCIPARUM*

2.2.1 *P. FALCIPARUM* CLONES

The 3D7 clone is chloroquine sensitive and was isolated in the Netherlands from the clinical isolate NF54, it is thought to originate from Africa (Myrick *et al.*, 2005). Dd2 was derived from the field isolate W2 following mefloquine selection (Guinet *et al.*, 1996). W2 is chloroquine resistant and originated in Thailand (Myrick *et al.*, 2005). AHE1 is a clone of Dd2 (Dd2^{attB}). The AHE1 clone has been transfected with an *AttB* docking site at the *cg6* locus of chromosome 7. The human dihydrofolate reductase (DHFR) selectable marker has also been transfected into this locus in order to maintain the *AttB* site via WR99210 selection (Nkrumah *et al.*, 2006).

2.2.2 *P. FALCIPARUM* CELL CULTURE MEDIUM

To make incomplete cell culture medium the following was added to a 500ml bottle of RPMI-1640 (Sigma): 18.75ml HEPES buffer (Sigma), 5ml filter sterilized 45% glucose solution (w/v in sdH₂O), 3ml filter sterilized 1M sodium hydroxide solution (sodium hydroxide in distilled water), 1.25ml gentamicin sulphate solution (Sigma), 5ml 200mM L-glutamine solution (Sigma) and 0.5ml of 1000X hypoxanthine solution (0.1M solution in 1M sodium hydroxide) were added. To make into complete cell culture medium 20ml of pooled human serum and 20ml of 5% Albumax II (w/v in sdH₂O) was added to each 500ml bottle (4% human serum, 0.2% albumax II).

2.2.3 HUMAN RED BLOOD CELL PREPARATION

Type O Rhesus positive leukocyte depleted red blood cells (RBC) (National Blood and Transfusion Service) were stored into 50ml aliquots at 4°C for 3-4 weeks. Prior to use the 50ml aliquot was centrifuged for 15 minutes at 3500rpm 4°C and the serum fraction removed

(pooled for future complete cell culture medium). The RBCs were then washed by the addition of an equal volume of incomplete media and collected by centrifugation for 10 minutes at 3500rpm at RT. The wash step was repeated twice more. The RBCs were re-suspended in an equal volume of incomplete media to give a 50% haematocrit and stored at 4°C for up to 10 days.

2.2.4 MAINTENANCE OF *P. FALCIPARUM* BY CONTINUOUS CELL CULTURE

P. falciparum infected RBC (iRBC) were maintained at 37°C in complete growth media at 1-2% haematocrit (HCT) in continuous culture (Trager and Jensen, 1976; Freese *et al.*, 1988). Life cycle staging and parasitaemia were established following methanol fixation of thin blood smears using 10% Giemsa staining. The slides were assessed by light microscopy (X1000 oil immersion lens) and appropriate volumes of prepared RBCs (50% HCT) and complete cell culture medium were added as required to maintain cultures at a 2-3% parasitaemia and 1-2% HCT. The cultures were gassed after medium changes to maintain an atmosphere of 1% oxygen, 3% carbon dioxide and 96% nitrogen (BOC).

2.2.5 CULTURE SYNCHRONISATION

Culture synchronisation was achieved via sorbitol lysis over several life-cycles (Lambros and Vanderberg, 1979). A predominantly ring stage culture was collected by centrifugation (5 mins, 1400rpm, RT), the supernatant removed and a X10 pellet volume of pre-warmed (37°C) 5% sorbitol w/v solution (Fluka) added. The culture was incubated for 5-10 minutes in a 37°C water bath, centrifuged (5 mins, 1400rpm, RT). The supernatant was removed and the pellet washed in complete cell culture medium and re-centrifuged (5 mins, 3000rpm, RT). After supernatant removal the culture pellet was re-suspended in an appropriate volume of complete cell culture medium, gassed and returned to the 37°C incubator.

2.2.6 TRANSFECTION

Dd2^{attB} ring stage iRBCs were co-transfected with pINT and the luciferase cassette within the pDCAttP plasmid using the direct electroporation technique as fully described by (Hasenkamp *et al.*, 2012a). Briefly, 200µl of 2 x cytomix, 100µl of packed iRBC (6-10% ring stage parasitaemia), 40µg pINT and 40µg of luciferase construct (made up with sdH₂O to a total volume of 400µl) were placed in electroporation cuvettes (4mm, chilled, BioRad) and subjected to a 0.31KV and 950 µF pulse using a GenePulser II electroporator (Biorad). The iRBCs were then immediately transferred to 10ml of pre-warmed cell culture medium, gassed and returned to the incubator (37°C).

The cell culture medium was changed daily for the first 8-10 days. From the second day after transfection, 100µl of fresh RBCs (50% HCT) were added every 2 days and 100µg/ml G418 sulphate (Sigma, UK) and 2.5µg/ml Blastocidin S hydrochloride (Invitrogen, UK), were added to the culture medium. After a further two weeks, only the 2.5µg/ml Blastocidin S hydrochloride was added to the cell culture medium to facilitate curing of the pINT plasmid. All cultures were maintained until parasites were detected via methanol fixation of thin blood smears followed by Giemsa staining. At this point, typically 28-35 days after transfection, the Blastocidin S hydrochloride selection was removed for two weeks to facilitate curing of un-integrated pDCAttP plasmid.

2.3 DNA METHODS

2.3.1 *P. FALCIPARUM* gDNA ISOLATION

Adapted from the standard phenol chloroform extraction protocol (Wong *et al.*, 2011). Note: all centrifugation steps are carried out for 5mins, 3000rpm at RT unless otherwise stated.

Approximately 1×10^9 trophozoite/schizont stage iRBCs were collected by centrifugation. The iRBC pellet was washed in 3ml of chilled PBS and collected by centrifugation. A 15X volume of PBS/0.1% saponin was added and gently mixed to lyse the iRBCs. After 2 mins, the released parasites were collected by centrifugation (5mins, 4000rpm, 4°C). 1840µl of DNA extraction re-suspension buffer (1200µl sdH₂O, 200µl TN9 [500mM Tris HCl pH9, 2M NaCl], 400µl 0.5M EDTA and 40µl 50mg/ml⁻¹ Proteinase K) was added to the parasite pellet and gently re-suspended by pipetting through a large bore tip. 200µl of SDS was added, mixed by inversion and flicking, and the tube incubated overnight at 50°C. 1ml of phenol/chloroform/isoamyl alcohol (25:24:1, Sigma) was added and the tube incubated whilst rotating for 60 mins at RT. The aqueous and phenol phases were separated by centrifugation. The upper aqueous phase was transferred to a clean microcentrifuge tube and the phenol extraction repeated. A 1/30th volume of NaOAc pH5 and 0.6 volumes of isopropanol (Sigma) was added to the aqueous phase and mixed by gentle inversion. Precipitated gDNA was spooled using a bacterial streaking loop and re-suspended in 100-200µl (depending on the size of the spooled DNA pellet) T₁₀E₁ (10mM Tris HCl, 1Mm EDTA, pH8.0). The gDNA was stored at 4°C until required.

2.3.2 MINIPREPARATION OF PLASMID DNA

Minipreparation of plasmid DNA from *Escherichia coli* was carried out using the Wizard Plus SV Miniprep DNA Purification System (Promega) in accordance with the manufacturer's recommended protocol. Briefly, 1.5ml of an overnight *E.coli* culture was centrifuged at 10,000g for 5 minutes to collect the cells. The cells were then suspended in 250µl Cell Re-suspension Solution and lysed by the sequential addition of 250µl Cell Lysis solution (mixed by inversion) followed by 10µl Alkaline Protease Solution (mixed by inversion). 350µl Neutralization Solution was then added (mixed by inversion) and the lysate cleared by centrifuging at 13,000g for 10 minutes. The supplied spin columns were placed on a Vacuum Manifold (Promega), the cleared lysate added and the vacuum applied to pull the liquid

through the columns. 750µl and 250µl aliquots of Wash Solution were also run through each column and the Vacuum Manifold run for a further 10 minutes to ensure that the columns were dry. Columns were then centrifuged at 13,000g for 2 minutes and placed in sterile 1.5ml microcentrifuge tubes. The plasmid DNA was eluted via the addition of 40µl Nuclease-Free Water and a 1 minute centrifuge at 13,000g.

2.3.3 MAXIPREPARATION OF PLASMID DNA

Maxipreparation of plasmid DNA from *E. coli* was carried out using the Plasmid Maxipreparation Kit (Qiagen, UK) in accordance with the manufacturer's recommended protocol. Briefly, 200ml of overnight culture was separated into 50ml Falcon tubes and the cells collected at RT by centrifugation (6,000g for 15mins). The *E.coli* pellets were suspended in a total of 3ml of P1 buffer and transferred to one 50ml Falcon tube. 10ml of P2 buffer was added and the tube mixed by inversion and allowed to incubate at RT for 5 minutes. 10ml of chilled P3 buffer was added, mixed by inversion and poured into the QIA filter cartridge. After 10 minutes incubation the supernatant was forced through the plunger into an equilibrated tip. The tip was then washed twice with 30ml of QC buffer and eluted with 15ml of QF buffer. 10.5ml of isopropanol was added, mixed and centrifuged (4°C, 30 minutes at 15,000g). The plasmid DNA was washed with 5ml of 70% ethanol and centrifuged (4°C, 10 minutes at 15,000g), air dried and re-suspended in 300-500µl of sdH₂O depending on the size of the pellet.

2.3.4 SPECTROSCOPIC ANALYSIS OF DNA

The absorbance of DNA and RNA samples were measured at 260 and 280nm using a Nanodrop 1000 Spectrophotometer (Thermo Scientific). These data were used to evaluate the quality (A^{260}/A^{280}) and quantity (A^{260}) of nucleic acids.

2.3.5 POLYMERASE CHAIN REACTION (PCR)

PCR was carried out using a Herculase® II Fusion DNA polymerase kit (Stratagene). The reaction mixture contained 16µl sdH₂O, 5µl 5X Herculase® II reaction buffer (1 X Mg²⁺ concentration 2mM), 1µl dNTP mix (40nM), 0.5µl DNA template (3D7 gDNA, *ca.* 100ng/ml), 0.5µl Herculase® fusion DNA polymerase and 1µl each of the forward and reverse primer (100pmol/µl). The PCR was carried out in a 0.2ml thin walled PCR tube (StarLab) in an Eppendorf Mastercycler Personal thermocycler (95°C/5 min [95°C/1 min, 48-52°C/1 min, 68°C/1 min] X 29, 68°C/10 min). The labelled tubes were then stored at -20°C until required.

Alternatively, the AccuPrime™ Pfx kit (Invitrogen) was utilized according to manufacturer's recommendation. This polymerase was chosen when cloning DNA fragments for transfection plasmids due to its fidelity as a result of its enhanced proofreading activity. In some instances Mg₂SO₄ concentration were varied (0.5mM, 1mM, 1.5mM, 2.0mM) to optimise results. The same cycling conditions as above were utilized.

If required, PCR products were cleaned up for restriction/ligation reactions using a commercial kit (Wizard® SV Gel and PCR clean-up system) according to the manufacturer's instructions.

2.3.6 GEL ELECTROPHORESIS AND IMAGING

Agarose gels (0.8-2%) were made using electrophoresis grade agarose (Invitrogen) and 1X TAE electrophoresis buffer. The appropriate weight of agarose powder was added to an appropriate volume of TAE and heated (1-2mins, 800W, microwave oven) until melted, then allowed to cool to approximately 50°C. An appropriate volume of Ethidium Bromide (EtBr) was then added (~5µl of 10mg/ml EtBr per 100ml of gel). The cast, comb and tank were all washed well with sdH₂O, assembled and the melted agarose poured into the cast. The gel was allowed to set at room temperature. When cool, the gel was placed in the tank with 1X TAE

buffer. The gels were then loaded with the samples and an appropriate DNA marker (Invitrogen) and run between 80 and 100V (depending on gel size) for *ca.* 1 hour. Size fractionated nucleic acids were visualised and photographed using the Gene Genius Bioimaging System (Syngene) and Gene Snap Software (Syngene). Where required, images were captured with a ruler for the subsequent evaluation of RNA/DNA standards migration.

2.3.7 TOPO CLONING

1µl - 4µl of fresh PCR product was combined with 1µl salt solution (1.2M NaCl, 0.06M MgCl₂) and made up to a total volume of 5µl with sdH₂O. 1µl of pCR®-Blunt II-TOPO® vector was mixed to this gently and incubated at room temperature (20-23°C) for 5 mins. The reaction mix was cooled on ice and was used immediately for bacterial transformation. Both Zero Blunt® TOPO® Cloning Kit (Invitrogen, UK) and TOPO TA Cloning® Kit for Sequencing (Invitrogen, UK) were used according to manufacturer's instruction.

2.3.7.1 BACTERIAL TRANSFORMATION

2µl of the TOPO® cloning reaction mix was added to a vial of 50µl of *One Shot® Chemically Competent *E. coli* (Invitrogen, UK). The vial was mixed gently and incubated on ice for 30 mins. The *E. coli* were heat-shocked (42°C) by placing the vial in a water bath for 45 secs and then immediately placed onto ice. 250µl of SOC medium was added and the *E. coli* incubated horizontally for 1hr at 37°C whilst shaking (200rpm). 50µl and 200µl volumes of the transformed *E. coli* were spread onto pre-warmed ampicillin-selection (50µg/ml) LB plates and incubated (inverted) at 37°C overnight.

*Note: One Shot® TOP10 *E. coli* cells contain the following features: *hsdR* and *mrcA* for efficient transformation of unmethylated and methylated DNA from PCR and genomic amplification respectively. *lacZΔM15* for blue/white colour screening and *recA1* to reduce the incidence of nonspecific cloned DNA recombination. **Genotype:** F- *mcrA* Δ(*mrr-hsdRMS-mcrBC*) Φ80*lacZΔM15* Δ *lacX74 recA1 araD139* Δ(*araleu*)7697 *galU galK rpsL* (StrR) *endA1 nupG*

2.3.8 RESTRICTION DIGEST

Typically, 3µl of miniprep plasmid DNA was combined with 1.5µl 10 X One-Phor-All buffer (Pharmacia), 1.5µl BSA (10mg/ml), 0.5µl appropriate restriction endonuclease(s) and 8.5µl dsH₂O and the samples were placed in a water bath at appropriate temperature (4hrs-overnight). 3µl loading buffer was then added and the samples run on a 1-1.5% agarose/TAE gel at 100V for 30-60 minutes. Where scaling was required to isolate DNA fragments for sub-cloning, a maximum of 0.5µg of plasmid DNA/10µl of restriction volume was used.

2.3.9 QUANTITATIVE PCR

To provide an indication of the relative copy number of the different pDC*AttP*-luciferase constructs in each transfectant, quantitative PCR was performed on a StepOnePlus RealTime PCR system (Applied Biosystems) with Power SYBR Green Mastermix (Applied Biosystems). *Luc* and seryl-tRNA synthetase (PF07_0073, PF3D7_0717700) oligonucleotide primers (1 µM), 50ng of genomic DNA template and cycling conditions of 95°C for 10 mins, followed by 40 cycles of 95°C for 15 s and 60°C for 60 s. Using the 2^{-ΔΔCt} method of relative quantitation, the relative copy number of *luc* (all comparisons made against the FL transfectants), as compared to that of the endogenous single copy seryl tRNA synthetase gene, were determined in three independent experiments.

2.3.10 SEQUENCING

Samples of plasmid DNA (20µl with c 5-10µg DNA in 6mM Tris HCl pH7.8) were provided to a commercial company (EuroFins Mwg, Germany) for sequencing. The returned sequence files (FASTA format) were aligned with the corresponding *P. falciparum* genome sequence from PlasmoDB (<http://www.plasmodb.org>). Sequences alignment was carried out using NClustalW2 available online from the European Bioinformatic Institute (<http://srs.ebi.ac.uk>).

2.4 RNA METHODS

2.4.1 *P. FALCIPARUM* RNA EXTRACTION

Total RNA was extracted, size-fractionated and blotted according to the protocol described by (Kyes *et al.*, 2000). In brief, a 0.5-1ml iRBC pellet was solubilised in 10-20 pellet volumes of pre-warmed (37°C) TRIzol (Invitrogen, UK) and incubated for 5 mins at 37°C. Chloroform (0.2 X TRIzol volume) was added and mixed via vigorous shaking and allowed to stand at RT for 2-3 minutes. The tube was then centrifuged (30 mins, 3000rpm, 4°C) and the aqueous upper layer containing the RNA was removed with a wide bore pipette, avoiding the interface which contains DNA. Isopropanol (0.5 X TRIzol volume) was mixed with the collected upper phase via inversion and incubated for 2 hrs at 4°C to precipitate the RNA. The isopropanol/aqueous phase was aliquoted into 1.5ml microfuge tubes and centrifuged (30 mins, 13-14krpm, 4°C). The supernatants were carefully removed and the RNA pellets washed with 0.5ml of cold (-20°C) 75% ethanol. The microcentrifuge tubes were then centrifuged for 5 mins (13-14krpm, 4°C), the supernatant removed and the tubes inverted and the RNA air-dried at RT. The RNA was dissolved in pre-heated (60-65°C) formamide (total of 100-200µl depending on the size of the RNA pellets), and snap-cooled on ice. The isolated RNA was stored at -80°C.

2.4.1.1 GEL ELECTROPHORESIS AND IMAGING

Briefly, 0.8-1.2% agarose gels were made using electrophoresis grade agarose (Invitrogen) in 1X TBE electrophoresis buffer. The tank, cast and comb were soaked overnight in 1% SDS, washed with sdH₂O then soaked for a further 10 mins with 3% hydrogen peroxide (H₂O₂) (Sigma Aldrich) and finally rinsed thoroughly with sdH₂O. The agarose was then prepared and cooled as outlined above, but no EtBr was added, instead freshly made 1M guanidine thiocyanate (Fluka) was prepared and added to give a final concentration of 5mM. The

agarose was then poured into the cast and allowed to set at room temperature. When cool the gel was placed in the tank with 1X TBE buffer and an appropriate RNA ladder (Invitrogen) loaded. The sample load volume per lane (5-10 μ g RNA) was made up to 15 μ l load volume with pre-warmed (37°C) formamide. The load samples were heated to 60°C for 5 mins and snap cooled on ice prior to loading. Gel loading buffer was added to blank lanes to monitor progress of the size fractionation and the gels were run at 110 volts for *ca.* 10-15mins and at 70-80 volts for a further 2 hours. The gels were post-stained in EtBr (100 μ l of 10mg/ml EtBr in 150-200ml 1 X TBE buffer) and photographed with a ruler to record the migration of RNA standards.

2.4.2 NORTHERN BLOTS

The agarose gel containing the size fractionated RNA was soaked in 7.5mM NaOH for 15 mins before capillary transfer to Hybond N⁺ in 7.5M NaOH overnight. The blot apparatus comprised 2 X 3MM Whatman paper wicks suspended over a glass plate into ~500ml of 7.5mM NaOH. The agarose gel was placed on top of the wick with 1 sheet of Hybond N⁺ membrane, 1 X wet and 2 X dry 3MM Whatman paper sheets placed on top. This was then covered with paper towels; 10X towels per stack, 2 stacks per layer with approximately 5-7 weighted layers. After overnight transfer, the Hybond N⁺ membrane was neutralized in 2X SSC and air dried. The filter was then placed on a UV light box and the position of RNA markers and rRNA clearly marked with a pencil. The membranes were then stored at RT until use.

2.4.2.1 PROBE HYBRIDISATION AND VISUALISATION

Blots were hybridised with DNA probes labelled using a random priming reaction with α^{32} P dATP. The Hybond N⁺ membranes were pre-hybridised in 10 ml of Church buffer at 55°C overnight. The DNA probes were labelled in accordance with the Random Primer labelling kit

(Amersham) protocol. Briefly, the DNA template (c. 100ng), random hexamer primer solution and sdH₂O were denatured for 5 mins at 98°C and allowed to cool to room temperature to facilitate primer annealing. Unlabelled dGTP, dCTP, dTTP, α³²P-labelled dATP and Klenow were then added and incubated at 37°C for 15-30 minutes. MicroSpin™ G-25 columns (Amersham) were prepared according to manufacturer's recommendation, and the labelling mix applied to the columns and the columns centrifuged (2 mins, 3000g, RT) to remove unincorporated nucleotides. The radiolabelled DNA was denatured for 2 minutes at 98°C, snap cooled on ice and then added to the Church buffer on the pre-hybridised blot. The radiolabelled DNA was hybridized to the Hybond N⁺ filter overnight at 55°C.

The radiolabelled Church buffer was removed and stored for disposal (according to Keele University radioactive waste protocol) and the Hybond N⁺ filter subject to a series of 30ml SSC/0.1% SDS washes of varying stringency (2X SSC, 0.5X SSC, 0.2X SSC, 0.1X SSC) to remove unbound radiolabelled probe. The the Hybond N⁺ filter was wrapped in clingfilm (Saranwrap) and placed in a film cassette with a blank phosphor screen for between 2-3 hours and overnight to expose. The images were recorded via a Cyclone Phosphoimager (Packard) and visualised using OptiQuant image analysis software.

2.4.3 5' RAPID AMPLIFICATION OF cDNA ENDS (RACE)

5' RACE was performed using the materials and protocols supplied with the 5' RACE kit (Invitrogen).

2.4.3.1 FIRST STRAND SYNTHESIS

To a 0.5ml PCR tube, 1-5µg of total RNA, ~10-25ng (2.5pmoles) of antisense gene-specific primer (GSP1) and sufficient DEPC-treated water to make a final volume of 15.5µl were mixed. This was incubated at 70°C for 10 mins then chilled on ice for a further minute. A brief centrifugation allowed product collection then, in order, the following were added to the

tube: 2.5µl of 10X PCR buffer, 2.5µl of 25mM MgCl₂, 1µl of 10mM dNTP mix and 2.5µl of 0.1M DTT. After mixing gently this was incubated at 42°C for 1 minute and then 1µl of SuperScript™ II RT was added, mixed gently and incubated at 42°C for 50mins. The reaction was terminated by incubating for 15mins at 70°C. After a brief centrifuge the reaction mix was placed at 37°C and 1µl of RNase added, mixed and incubated for 30 mins then placed on ice.

2.4.3.2 SNAP COLUMN PURIFICATION OF cDNA

120µl of binding solution (6M NaI) was added to the first strand reaction and transferred to a SNAP column. This was centrifuged (13,000g) for 20 secs. Next, 0.4ml of cold (4°C) 1X wash buffer was added to the spin cartridge which was re-centrifuged as above. This step was repeated three times. The cartridge was then washed with 400µl cold (4°C) 70% ethanol and centrifuged, as above, twice. This was followed by a final centrifuge (3,000g) for 1 minute. The cartridge was then placed in a recovery microcentrifuge tube and the cDNA eluted with the addition of 50µl of pre-warmed (65°C) sdH₂O and centrifuged (13,000g) for 20 secs.

2.4.3.3 TdT TAILING OF cDNA

6.5µl DEPC-treated water, 5.0µl of 5X tailing buffer and 2.5µl of 2mM dCTP were added to 10µl of the SNAP purified cDNA. This was incubated at 94°C for 2-3 mins then chilled on ice (1min). A brief centrifugation collected the contents, which were then placed back on ice. To this, 1µl of TdT was added, mixed and incubated at 37°C for 10 mins. The TdT was inactivated by placing at 65°C for 10 mins. To a 0.5µl microcentrifuge tube on ice, 31.5µl of sdH₂O, 5.0µl of 10X PCR buffer, 3.0µl of 25mM MgCl₂, 1.0µl of 10mM dNTP mix, 2.0µl of gene-specific primer 2 (GSP2) (10µM), 2.0µl of Abridged Anchor Primer (10µM) and 5.0µl of the dC-tailed cDNA and 0.5µl of Taq DNA polymerase (5 units/µl) was added prior to gentle mixing. The tube was then transferred directly for PCR (thermocycler pre-equilibrated to 94°C) and 30-40 cycles of

amplification were undertaken. PCR conditions: 94°C for 1min, 55°C for 1 min, 72°C for 1min (X30-40) followed by 72°C for 7mins and a 5°C hold temperature.

2.4.3.4 NESTED AMPLIFICATION

5µl of the primary PCR reaction was then diluted with 495µl of TE buffer. To a 0.5ml microcentrifuge tube on ice; 33.5µl of sdH₂O, 5.0µl of 500mM KCl, 3.0µl of 25mM MgCl₂, 1.0µl 10mM of dNTP mix, 1.0µl of nested GSP (10µM), 1.0µl of Abridged Universal Amplification Primer (AUAP)/Universal Amplification Primer (UAP) (10µM) and 5.0µl diluted PCR product were added. 0.5µl of Taq DNA polymerase was added prior to mixing and the tube immediately placed in the pre-heated thermocycler. To enhance the amplification of the nested product, various final concentrations of Mg²⁺ were used at this stage. Any PCR products generated were cloned using the TOPO cloning protocol (See section 1.4.7)

2.4.4 3' RACE

3' RACE was performed using the materials and protocols supplied with the GeneRacer™ Kit (Invitrogen).

2.4.4.1 FISRT STRAND SYNTHESIS FROM OLIGO-dT

Briefly, 10µl of total RNA (2µg in 10µl), 1µl of primer (Oligo-dT primer supplied with kit), 1µl of dNTP (10mM) and 1µl of sdH₂O were incubated for 5 mins at 65°C, chilled on ice (1 minute), then centrifuged briefly to collect reaction. To this, 4µl of 5X first strand buffer (250mM Tris-HCL, 374mM KCL and 15mM MgCl₂), 1µl of DTT (0.1M), 1µl of RNase Out (40U/µl) and 1µl of Superscript III reverse transcriptase (200U/µl) were added, centrifuged briefly to collect and incubated at 50°C for 30 to 60 minutes. The reverse transcription reaction was inactivated by incubating for 15 minutes at 70°C before 1µl of RNase H (2U/µl) was added and the reaction mixture incubated for a further 20 minutes at 37°C. The prepared

cDNA was stored at -20°C until required. A control reaction, omitting the Superscript III reverse transcriptase (-RT) was also prepared.

2.4.4.2 3' RACE PCR

A PCR reaction master mix was prepared that contained; 10µl of 10X AccuPrime™ Pfx reaction mix (contains dNTPs), 3µl of the gene specific primer (10ng/ml), 2µl of the GeneRacer™ 3' primer, 2µl of AccuPrime™ DNA Polymerase and 79µl of sdH₂O. One half of the reaction mix was used with the first strand cDNA (varying the Mg²⁺ to optimise the amplification), with the second half being used with the -RT control material. If a further round of nested amplification was required, an internal GeneRacer™ 3' primer was available. Any PCR products generated were cloned using the TOPO cloning protocol (See section 1.4.7)

2.5 LUCIFERASE ASSAY

The luciferase assay used here followed the format of the revised single-step lysis protocol described by (Hasenkamp *et al.*, 2012b). All parasite cultures were synchronized and matched for staging and parasitaemia. Parasite cultures at 1% parasitaemia and 2% HCT were sampled from early rings (T1, 6-12hrs), late rings (T2, 12-18hrs), early trophozoites (T3, 18-26hrs), mature trophozoites (T4, 26-34hrs) and schizonts (T5, 36-42hrs) (200µl harvested per sample). The culture flasks were then gassed and returned to incubator (37°C). From the 200µl aliquot taken, 3 x 40µl samples were utilized for the luciferase assay, the remainder of the sample was used to create a thin blood smear. Each of the 40µl parasite culture samples were added to 10µl of passive lysis buffer (Promega, UK) in a 96-well white multiplate (Greiner, UK) and mixed well. A 50µl aliquot of luciferase assay buffer (Promega, UK) was then added to each well. Bioluminescence - (relative light units, RLU) was measured using a GloMax Multi Detection System (Promega, UK) for 2 seconds. Data analysis: parasitaemia was normalized against 1 x 10⁶ mature trophozoite stage RLU. Statistical

analysis from 3 independent experiments (total n=9) was done using ANOVA with Tukey's post-test for significance.

2.6 EXPERIMENTAL OLIGONUCLEOTIDES

The following tables provide the name(s) and sequences of oligonucleotides used in this study. These are organised according the project and/or methodology for which they were employed. Where relevant the sizes of the expected PCR product or notes relating to their application, and/or inclusion of restriction enzyme sites (sites underlined in oligonucleotide sequence) are added.

Table 2-1 Oligonucleotides for northern blot analyses

Name	Unique ID	Sequence	Size	Notes
KR1107PFA0545cF	NBP1	TGAAAGGGTTTTGTATTAGGAGG	591bp	Probe for Northern blot
KR1107PFA0545cR	NBP2	ATCTTACACCAAAGGTAGATGTGG		with NBP1
KR1107PFB0895cF	NBP3	CCATGGGGAATATCTAGAAGATGG	622bp	Probe for Northern blot
KR1107PFB0895cR	NBP4	ATTATGGAATATCCACCTGTAGCC		with NBP3
KR1107PFF1225cF	NBP5	AAAGTTGGGTTATATGGTTGGGG	794bp	Probe for Northern blot
KR1107PFF1225cR	NBP6	CTCCGTGTGATCTGTGCTAATGG		with NBP5
KR1107PFI0235wF	NBP9	GTGTAAGTCTGGGGTAGTTCAGC	683bp	Probe for Northern blot
KR1107PFI0235wR	NBP10	GCACATCACTAACATCACTTATGG		with NBP9
KR1107PFI0530cF	NBP11	AGTGGAATTCACATGAGAAGGGG	818bp	Probe for Northern blot
KR1107PFI0530cR	NBP12	TATATCTGGTACATATGCTACACC		with NBP11
KR1107PF10_0154F	NBP13	TATTCAGATACCTGAAGGAAGAGC	592bp	Probe for Northern blot
KR1107PF10_0154R	NBP14	CGGGCCAACAAAAGGGTACTTGG		with NBP13
KR1107PF10_0165F	NBP15	ACCAAATGTATACGATGAGATAGG	570bp	Probe for Northern blot
KR1107PF10_0165R	NBP16	ATTAGCTTCATATACAATACCACC		with NBP15
KR1107PF11_0282F	NBP17	AACACATCATGAAGGTGATAGTGG	428bp	Probe for Northern blot
KR1107PF11_0282R	NBP18	CCTTCTCCTCTGGAAGTTTCATCC		with NBP17
KR1107PF13_0291F	NBP19	GAAACATGTATTAATGTGGGAACC	582bp	Probe for Northern blot
KR1107PF13_0291R	NBP20	TCCTGTAACATAAATACGATCTCC		with NBP19
KR1107PFL1285cF	NBP21	CGGTGATAAAAGAAAAGTCGACCG	525bp	Probe for Northern blot
KR1107PFL1285cR	NBP22	GTCAAATGAGATGGATCTTGTAGG		with NBP21
KR1107PFL1915wF	NBP23	TATTGGGAACACTGATGTGAAAGG	565bp	Probe for Northern blot

KR1107PFL1915wR	NBP24	GCGTTCCTCTCTTATTAATTGGGC		with NBP23
KR1107PFL2005wF	NBP25	CCTGGTACGGGTTAAAACGACAAGC	760bp	Probe for Northern blot
KR1107PFL2005wR	NBP26	CCTGTTGCCATGGTATTACATGCC		with NBP25
KR1107PF13_0328F	NBP27	GGGAATCATGTATCTTTGGTTAGC	631bp	Probe for Northern blot
KR1107PF13_0328R	NBP28	GGTATCTGAATCAGGGGAGGTATC		with NBP27
KR1107PF14_0053F	NBP29	TGTTGCTAATGCATGTGTTGAAGG	463bp	Probe for Northern blot
KR1107PF14_0053R	NBP30	CCTTTCGTTGAGCCATGACTCCTG		with NBP29
KR1107PF14_0148F	NBP31	TCCTATAGGAGTAAAAATACCTCC	412bp	Probe for Northern blot
KR1107PF14_0148R	NBP32	GGGGTAGCTCCCATTTGATGGGCG		with NBP31
KR1107PF14_0177F	NBP33	AGAAAGACAAGCCGAGCTAGAAGG	755bp	Probe for Northern blot
KR1107PF14_0177R	NBP34	GGAGATATTCTATATGCTTCTACC		with NB33
KR1107PF14_0254F	NBP35	TCGACCAGTAATAGTCGATCATGG	544bp	Probe for Northern blot
KR1107PF14_0254R	NBP36	CATGTCTGTTAATAACCCCTTAC		with NBP35
KR1107PFB0840wF	NBP37	TCCGTGGGTTGAAAAGTACCGACC	620bp	Probe for Northern blot
KR1107PFB0840wR	NBP38	CTGCTCTTCTAAATCACCTTCGG		with NB37
KR1107PFE1345cF	NBP39	CAACCGGACAATGCCTAGATGGG	697bp	Probe for Northern blot
KR1107PFE1345cR	NBP40	CGTAGTGTGTATACCTGCCTTAGC		with NBP39
KR1107PFL1655cF	NBP41	CGTTATATGTAATACATATCCAG	638bp	Probe for Northern blot
KR1107PFL1655cR	NBP42	CTGAATTGGATATTATGGAATCGC		with NB41
KR1107PF11_0117F	NBP43	CCTGGTGGAGGGAAAAGTACTCGC	621bp	Probe for Northern blot
KR1107PF11_0117R	NBP44	CAGCTACTGATAAACTTTGATGAG		with NB43
KR1107PF14_0362F	NBP47	GGAGGAAGAATATGGAAGCAGC	746bp	Probe for Northern blot
KR1107PF14_0362R	NBP48	CTCTCAGTTAGTACGTCAAAGGTG		with NBP47
KR1107PFC0340wF	NBP49	GACCGTTATAGTAAAAGGTATTGG	611bp	Probe for Northern blot
KR1107PFC0340wR	NBP50	TCCTAGAGTATCTGGAGACGTGGG		with NBP49
KR1107PF14_0602F	NBP51	TGAAGGACATAAAACAGGCAGACG	559bp	Probe for Northern blot
KR1107PF14_0602R	NBP52	ACCATTATTCATTTTCATATCCC		with NB51
KRPFA0285cF	NBP53	CGAAATTGATGACACAGACG	688bp	Probe for Northern blot
KRPFA0285cR	NBP54	CCATCCAGTACTCATTGGG		with NBP53
KRPF110424F	NBP57	GGACCTACATGATAATGC	518bp	Probe for Northern blot
KRPF110424R	NBP58	CAAAAAGGGTATCTAACC		with NBP57
KRPF110425F	NBP59	GATATGAACATTTGTACGG	429bp	Probe for Northern blot
KRPF110425R	NBP60	CCATATTGTTACATCCCAAC		with NBP59
KRPF130199F	NBP65	GATTCAAAAATATAAAACAGG	441bp	Probe for Northern blot
KRPF130199R	NBP66	CCATATTTTTATTACTTAGTTCG		with NBP65
KRPF130200F	NBP67	GGAAATAATTGTAAAGAGAGG	539bp	Probe for Northern blot
KRPF130200R	NBP68	CCAGCATCATATGATTTGTC		with NBP67
KRPF070064F	NBP69	CCAAAGGCAGCTTATATGCC	554bp	Probe for Northern blot
KRPF070064R	NBP70	GGGTGTTAAATAATTGAGTGC		with NBP69

KRPF070065F	NBP71	GGAAATAAAATAGATGGAACG	579bp	Probe for Northern blot
KRPF070065R	NBP72	CGGCTATTGAATATTTAGG		with NB71
KRPF10725cF	NBP73	CCTGTTGTGGATTATATGGG	444bp	Probe for Northern blot
KRPF10725cR	NBP74	CCTTCGATGTTCTCCAG		with NBP73
KRPF140149F	NBP75	CGTGCAAAGAAATAATGC	525bp	Probe for Northern blot
KRPF140149R	NBP76	CCACAGAATAACTTTTGCCG		with NBP75
KRPF140295F	NBP77	GGTTTAACTACAGGACAAGC	610bp	Probe for Northern blot
KRPF140295R	NBP78	GCTTGTCTTGATTCTGACC		with NBP77
KRPF0595wF	NBP79	CGAATTTAACCAATATGAGG	620bp	Probe for Northern blot
KRPF0595wR	NBP80	CCCGTTTGTCATTAAGGGG		with NBP79
PFI0540w-FOR	NBP81	ATATTACTTCTCAGTAGATATGG		Probe for Northern blot
PFI0540w-REV	NBP82	TAATCCACCCTTCACTATTTTGC		with NBP81
M15077-FOR	P19	CAAATCACAGAATCGCGTATGC		Luc Probe for Northern blot
M15077-REV	P20	TCGAAATCCACATATCAAATATCC	563bp	Luc with P19
PFD0660w-FOR	P21	GACTTATGCATATGAGAAATTCG		Probe for Northern blot
PFD0660w-REV	P22	TACTTGAAAATCGTAACCATGC	544bp	with P21
PFD0660w-FOR	NBP83	GACTTATGCATATGAGAAATTCG	544bp	Probe for Northern blot
PFD0660w-REV	NBP84	TACTTGAAAATCGTAACCATGC		with NBP83

Table 2-2 Oligonucleotides for PFD0660w 5' UTR deletion study

name	Unique ID	Sequence		Size	Notes
PFD0660w- <i>Apal</i> -for	P1	GTGGGCCACCATACACAAATAAACATAGACAC	<i>Apal</i>		cloning 5'
PFD0660w- <i>HindIII</i> -full- rev	P2	TCAAGCTTTTTTACTCATATTTTTTTTTTATATATATTC TCACACAC	<i>HindIII</i>	1286bp	cloning with P1
PFD0660w- <i>HindIII</i> - Del1- <i>rev</i>	P3	TCAAGCTTTGTGTAATAATAACACATCTGAAC	<i>HindIII</i>	1061bp	cloning with P1
PFD0660w- <i>HindIII</i> - Del2- <i>rev</i>	P4	TCAAGCTTATATTTAACGCACTAAATCCTCACC	<i>HindIII</i>	870bp	cloning with P1
PFD0660w- <i>HindIII</i> - Del3- <i>rev</i>	P5	TCAAGCTTATCATTATAATATTATAGTGAAAACC	<i>HindIII</i>	625bp	cloning with P1
*PFD0660w- <i>HindIII</i> - Del4- <i>rev</i>	P31	CCAAGCTTCTCATTTTTATTTATATCTTTCTCC	<i>HindIII</i>	343bp	cloning with P1
Luc-ORF- <i>HindIII</i> -for	P6	CTAAGCTTATGGAAGACGCCAAAAACATAAAGAAAGG	<i>HindIII</i>		cloning <i>luc</i>
Luc-ORF- <i>Sall</i> - <i>rev</i>	P7	TGCGTCGACTTACAATTTGGACTTCCGCCCTTCTTGG	<i>Sall</i>	1653bp	cloning with P6
PFD0660w- <i>HindIII</i> - <i>Sall</i> - 3UTR-for	P8	TCAAGCTTACTGTCGACTAATCCCTTTGAATTATG	<i>HindIII</i> / <i>Sall</i>		cloning 3'
PFD0660w- <i>PstI</i> -3UTR- rev	P9	ACGCTGCAGACGAATGTTATAATGATTACAGG	<i>PstI</i>	674bp	cloning with P8

* denotes work completed by the Horrocks lab after completion of my practical work

Table 2-3 Oligonucleotides for genomic integration into *AttB* locus

Name	Unique ID	Sequence	Size	Notes
pcg6-ORF-upstrea	P10	ATGAACAAATACATAAGAGCGC		over attL (gDNA) integration
attP-downstream2	P11	TAAGGAGAAAATACCGCATCAGG	568bp	over attL (gDNA) with P10
Delta1-attP-upst	P12	TATATAAGGACATATTTATTAACC		over attP (plasmid) episomal
Delta1-attP-down	P13	ACGGCCAGTGAATTGTAATACG	201bp	over attP (plasmid) with P12

Table 2-4 Oligonucleotides for qPCR of luciferase copy number and expression

Name	Unique ID	Sequence	Size	Notes
Luc-q1ss	P15	GAAGCTATGAAACGATATGG		<i>Luc</i> qPCR
Luc-q1as	P16	TGCAACTCCGATAAATAACG	110 bp	<i>Luc</i> with P15
PF07_0073-qss	P17	TCAATTTGATAAAGTGGAAACA ATTC		<i>PF07_0073</i> qPCR
PF07_0073-qas	P18	GCGTTGTTTAAAGCTCCTGA	152bp	<i>PF07_0074</i> with P17

Table 2-5 Oligonucleotides for 5' and 3' RACE

Name	Unique ID	Sequence	Notes
3'TrophGSP1	P23	ATAATTGTGGTATTACATGG	GSP1 for <i>PF0660w</i> 3'RACE
3'TrophGSP2	P24	TGTAGTTTTAAGAGAATTTGG	GSP2 for <i>PF0660w</i> 3'RACE
5'TrophGSP1	P25	ATCATATCTACGATTCC	RTPCR for <i>PF0660w</i> 5'RACE
5'TrophGSP2	P26	TACGAATTTCTCATATGC	GSP1 for <i>PF0660w</i> 5'RACE
5'TrophGSP3	P27	ATATTTGTTTCGTTTAC	GSP2 for <i>PF0660w</i> 5'RACE
5'LucGSP1	P28	AAGAGAGTTTTCACTGC	RTPCR for <i>luc</i> 5'RACE
5'LucGSP2	P29	ATGTTACCTCGATATGTGC	GSP1 for <i>luc</i> 5'RACE
5'LucGSP3	P30	ATCCTTAGAGGATAGAATGG	GSP2 for <i>luc</i> 5'RACE

2.7 BIOINFORMATICS

2.7.1 INTERGENIC DISTANCE ANALYSES

Two algorithms were used for the analysis; `PlasmoDB.gff.sorter.pl` and `Intergenic.dist.pl`. These were designed by myself, Paul Horrocks and Richard Emes and were coded in PERL script by Richard Emes (Keele University and then Nottingham University). They were run by myself within a UNIX environment. The organism general file format (.gff) file (latest version at the time the work was carried out) was downloaded from PlasmoDB or EupathDB and parsed using `PlasmoDB.gff.sorter.pl` to extract the relevant information required for the analysis of intergenic distances (gene identifier, chromosome number, coordinates and gene orientation). The parsed .gff file was then run through `Intergenic.dist.pl` to give four intergenic region (IGR) files depending upon the orientation of the genes that flanked the IGR namely; Head to Head (HH), Head to Tail (HT), Tail to Head (TH) and Tail to Tail (TT). Each output file contained the unique identifiers for the genes flanking the intergenic regions and the size of this region in base pairs (bp). A workflow and example output files are illustrated in Figure 1. These scripts are available from website (<https://sites.google.com/site/emesbioinformatics/group-software>).

Figure 2-1 Schematic representing workflow to capture intergenic distances for HH, HT, TH and TT gene flanking regions

(overpage) The latest .gff files were downloaded from PlasmoDB or EupathDB (where a .gff file was unavailable the cDNA file was used in its stead). For each organism the relevant information was then parsed, or filtered, from the .gff or cDNA file using the Perl script `PlasmoDB.gff.sorter.pl` (written by Dr Richard Emes). This process gave a file containing the gene name, chromosome, exon start and end co-ordinates and the gene's orientation. A second Perl script `intergenic.dist.pl` (written by Dr Richard Emes) was then applied to the parsed file. This script sorted the data into four files according to the orientation of the genes adjacent to the intergenic space and reported the two adjacent gene names and the total intergenic distance between them in bps. Where required for further analyses, the HT and TH files were concatenated and the datasets re-termed Type A, B and C respectively (A = HH, B = HT and TH and C = TT).

Sample .gff file

Pf3D7_01_v3	ApiDB	gene	29510	37126	.	+	.	ID=Pf3D7_0100100;Name=Pf3D7_0100100;description=erythrocyte+membrane+protein+142C+PfEMP1+3
Pf3D7_01_v3	ApiDB	mRNA	29510	37126	.	+	.	ID=rna_Pf3D7_0100100-1;Name=Pf3D7_0100100-1;description=Pf3D7_0100100-1;size=7617;Parent=P
Pf3D7_01_v3	ApiDB	CDS	29510	34762	.	+	0	ID=cds_Pf3D7_0100100-1;Name=cds;description=.;size=5253;Parent=rna_Pf3D7_0100100-1
Pf3D7_01_v3	ApiDB	CDS	35888	37126	.	+	0	ID=cds_Pf3D7_0100100-1;Name=cds;description=.;size=1239;Parent=rna_Pf3D7_0100100-1
Pf3D7_01_v3	ApiDB	exon	29510	34762	.	+	.	ID=exon_Pf3D7_0100100-1;Name=exon;description=exon;size=5253;Parent=rna_Pf3D7_0100100-1
Pf3D7_01_v3	ApiDB	exon	35888	37126	.	+	.	ID=exon_Pf3D7_0100100-2;Name=exon;description=exon;size=1239;Parent=rna_Pf3D7_0100100-1



Sample parsed .gff file

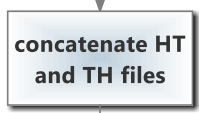
PFA0005w	MAL1	29733	37349	+
PFA0010c	MAL1	39205	40430	-
PFA0015c	MAL1	42590	46730	-
PFA0020w	MAL1	50586	51859	+
PFA0025c-1	MAL1	53392	53503	-
PFA0025c	MAL1	53392	53503	-
PFA0030c	MAL1	54001	55229	-



rename HH = A IGD file
rename TT = C IGD file

Sample HH file

PFA0015c	PFA0020w	3856
PFA0035c	PFA0040w	2889
PFA0055c	PFA0060w	1207
PFA0070c-1	PFA0075w-1	2173
PFA0100c	PFA0105w	1191
PFA0130c	PFA0135w	3104
PFA0170c	PFA0175w	1399



For the analysis of intergenic regions in the chromosome internal and subtelomeric regions, a PERL script `split.list.pl` (written by R. Emes) was used to separate information about types A, B and C intergenic distance files into different chromosomal compartments. This script used a list of unique gene identifiers (e.g. a list of genes determined to be within the subtelomeric region) to extract out the relevant information (e.g. from the type A intergenic distances file). Figure 2 illustrates the workflow used for this analysis.

To explore the correlation between intergenic distances and the temporal transcription of the flanking genes, the `split.list.pl` script was used to sequentially extract the types A, B and C intergenic distance files for each temporal window of transcription. The first step was to extract only those gene pairs where temporal intraerythrocytic schizogony microarray data (Bozdech *et al.*, 2003) was available for both genes. The remaining gene pairs were then parsed against groups of temporally co-transcribed genes, clustered into five groups as described by Jurgelenaite *et al.* (2009). In this way, for each of the types A, B and C intergenic distance files, a window of transcription during intraerythrocytic schizogony was annotated for each gene in the gene-pair list (see workflow and example data in Figure 3). Five windows of temporal transcription were used; 1, early rings (ER), 2, rings/early trophozoites (RET), 3, trophozoites/early schizonts (TES), 4, schizonts (S) and 5, constitutively transcribed (Jurgelenaite *et al.*, 2009). Filtering the final dataset in EXCEL allowed the number and sizes of intergenic regions subject to co-transcription to be readily identified.

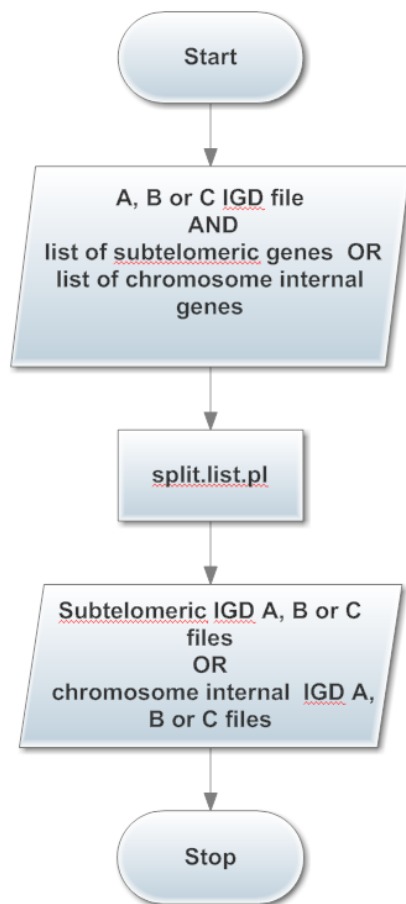
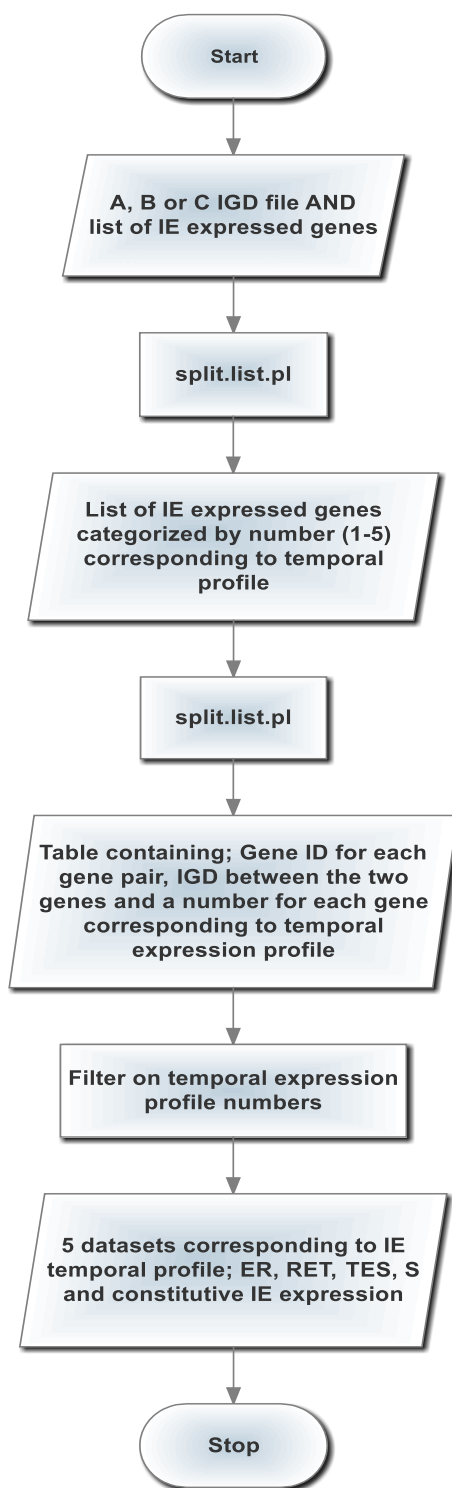


Figure 2-2 Schematic representing workflow to secure intergenic distances in different *P. falciparum* chromosome compartments.

Subtelomeric breakpoints were established for each *P. falciparum* chromosome (see chapter 3 main text). These breaks were based on loss of gene synteny between *P. falciparum*, and *P. vivax* and *P. knowlesi*. The gene names were then compiled into separate subtelomeric or chromosome internal files. Either the A, B or C IGD file and either the subtelomeric or chromosome internal files were used to create 6 separate files; A - subtelomeric, B - subtelomeric, C - subtelomeric, A - chromosomal internal, B - chromosomal internal and C - chromosomal internal using the Perl script `split.list.pl` (written by Dr Richard Emes). Each of these six IGD files contained the two adjacent gene names and the intergenic distance between them in bps.

Figure 2-3 Schematic representing work-flow to categorise temporal windows of transcription during intraerythrocytic schizogony for gene-flanking pairs intergenic regions.

(overpage) Initially the IGD A, B and C files were filtered, using `split.list.pl` (written by Dr Richard Emes), by the Bozdech *et al.*, 2003 IE dataset to create a dataset for which microarray data was available. Each A, B or C file was then filtered again, using `split.list.pl`, by the Jurgalenaite *et al.*, 2009 IE temporal expression profile gene lists. Each IE gene ID in this list had a number recorded against it (1-5) corresponding to its IE life-cycle stage expression profile; (1) ER (early ring), (2) RET (ring/early trophozoite), (3) TES (trophozoite/early schizont) and (4) schizont. The number 5 denoted constitutive IE expression. This gave a table containing the two adjacent gene IDs, the corresponding intergenic distance in bps and two numbers (1-5) corresponding to temporal expression profiles for each gene respectively. This table was filtered in Excel to give 5 temporal IE datasets.



Sample file used to create co-expression datasets

Gene left	Gene right	IGD	Expression gene left	Expression gene right
PF11_0503	PF11_0504	6098	1	1
PFF1530c	PFF1535w	1200	1	1
PF07_0005	PF07_0006	3620	1	1
PF08_0003	MAL8P1.4	4995	1	1
PFA0230c	PFA0235w	781	2	2
PFA0475c	PFA0480w	891	2	2
PF10_0086	PF10_0087	1068	2	2
PF10_0149	PF10_0150	1644	2	2
PF10_0209	PF10_0210	1418	2	2
PF10_0266	PF10_0267	1846	2	2
PF10_0277	PF10_0278	983	2	2
PF10_0299	PF10_0300	857	2	2
PF10_0322	PF10_0323	6786	2	2
PF10_0324	PF10_0325	944	2	2
PF10_0327	PF10_0328	1536	2	2

2.7.2 GENE SELECTION FOR NORTHERN BLOTS

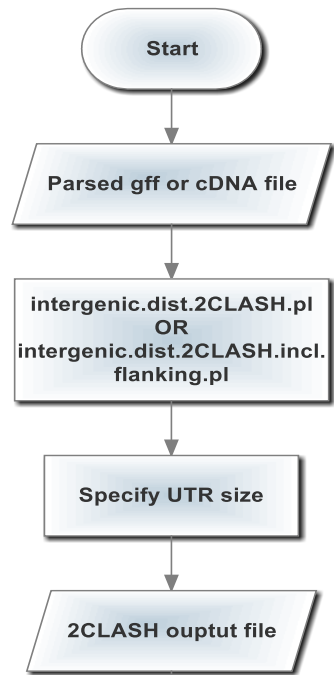
PubMed was searched using a combination of key terms; *Plasmodium, falciparum*, malaria, transcription, transcript, northern, expression. Publications identified were examined to provide gene identification information and whether northern blot data were presented. Northern blot data were accepted if the publication specifically described the transcript size or the northern blot (with markers) was shown. Northern blot data were rejected if the data were from a multi-gene family as specific annotation for an individual gene was not possible for many of these genes. The *de novo* Northern blot data reported in this study were carried out in the Horrocks laboratory; many by myself for this project, with a small number as part of other on-going studies. The first cohort of genes selected for northern blot analysis were trophozoite specific genes up-regulated during DNA replication (genes with similar PlasmoDB GO ontology and similar transcription profiles (Pearson coefficient of $R^2 > 0.95$)) as it was my original intention to study co-regulated genes. However, my research took an alternative path. The second cohort of genes selected for northern blot analysis, were selected in pairs on the basis of the same temporal IE expression and adjacent genomic location. Information relating to open reading frame size, orientation and number of exons were secured from PlasmoDB and the malaria IDC comparison database (<http://malaria.ucsf.edu/comparison/>) was used to identify the peak of temporal expression profiles and allocate life-cycle stages for transcription to each gene (Appendix D-2). Expressed Sequence Tag (EST) data were also secured from a combination of the Full-Malaria database (fullmal.ggc.jp), (Watanabe *et al.*, 2004) and dBEST, in each case securing the most distal EST coordinate available on either side of the ORF.

2.7.3 TRANSCRIPT APPORTIONMENT

Two algorithms, `intergenic.dist.2CLASH.pl` and `intergenic.dist.2CLASH.incl.flanking.pl`, were used to model 'transcript fit' over available intergenic space on a genome-wide scale. The algorithms for the analysis were designed by myself, Paul Horrocks and Richard Emes. These were coded into PERL script by Richard Emes and run by myself within a UNIX environment. The input file comprised the parsed organism .gff file (i.e. gene identifier, chromosome number, coordinates and orientation data provided) and allowed a variable user-specified fixed-length UTR length to be input for each analysis (e.g. 600 or 1800bp). A detailed description of the algorithms used is provided in chapter 5. The output data comprised an EXCEL spread sheet containing consecutive gene triplet identifiers with the corresponding IGR lengths, orientation of the flanking genes and a pass or fail value (0 or 1 respectively) for each 1% UTR 'fit' ratio over the central "gene of interest". The workflow and example data are shown in Figure 4. Using the available data for each fixed length UTR used over all genes in the genome (except the most distal gene on either end of the chromosome as no flanking gene is available), a cumulative fail rate for each transcript apportionment was calculated. The `intergenic.dist.2CLASH.pl` and `intergenic.dist.2CLASH.incl.flanking.pl` scripts are available from website (<https://sites.google.com/site/emesbioinformatics/group-software>).

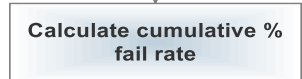
Figure 2-4 Schematic representing the workflow used to model transcript apportionment in *P. falciparum*.

(overpage) The parsed .gff or cDNA file was used as the input file for both; (1) `intergenic.dist.2CLASH.pl` or (2) `intergenic.dist.2CLASH.incl.flanking.pl` (both scripts were written by Dr Richard Emes). The UTR size was also an input parameter. (1) took each triplet of genes (right, central and left) and apportioned the specified UTR size over the available intergenic distance on either side of the central gene of the triplet in 1% increments. If the UTR 'fitted' at a specific 5'/3' % ratio it scored a 0. If it did not fit at a specific 5'/3' % ratio i.e. it overlapped with the upstream or downstream gene ORF it scored a 1. (2) was more stringent. This script apportioned the specified UTR size over the available intergenic distance of the central gene of triplet of genes as above but, in addition, it apportioned the specified UTR size (at 1% increments) over the intergenic space of the left and right gene IGR also. Hence, it did not just take into account whether there was an ORF 'clash' but also whether there was a 'clash' with the apportioned UTR to the right and left gene. The % cumulative fail rate was then calculated and graphed against % occupancy.



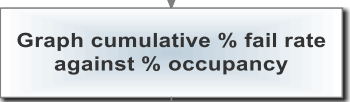
Sample 2CLASH raw output file

Left Gene	Left Gene (Start_End_strand)	inter_type	Current Gene	Current Gene (Start_Peptide len)	Coding len	inter_type	Right Gene	Right Gene (Start_Er)	Upstream	Downstream	Max Perce	Max Perce	100	99	98	97	96
MAL13P1.184	1491124 1494504 -	HZH	MAL13P1.185	1495455 1497436 +	305	915 T2H	MAL13P1.186	1498073 1501699 +	351	537	68	46	1	1	1	1	1
MAL7P1.167	1361270 1369591 -	HZH	MAL7P1.170	1376682 1377743 +	293	879 T2H	MAL7P1.171	1380751 1387189 +	7091	3008	100	100	0	0	0	0	0
MAL7P1.230	92135 93136 -	HZH	MAL7P1.228	106224 108526 +	661	1983 T2T	MAL7P1.229	110392 115696 -	13088	1866	100	100	0	0	0	0	0
PF07_0034	454668 455621 -	HZH	PF08_0034	458600 461696 +	424	1272 T2H	PF07_0035	463593 467339 +	2979	1898	100	100	0	0	0	0	0
MAL8P1.40	1023701 1024918 -	HZH	PF08_0034	1029824 1034596 +	1485	4455 T2H	MAL8P1.38	1034637 1037500 +	4906	51	100	4	0	0	0	0	0
MAL8P1.153	152929 160789 -	HZH	PF08_0131	166536 167198 +	220	660 T2T	PF08_0130	168189 171554 -	5747	991	100	71	0	0	0	0	0
PF10_0015	68690 68962 -	H2T	PF10_0016	70548 70820 -	90	270 H2T	PF10_0017	72935 73918 -	2115	1586	100	100	0	0	0	0	0
PF10_0083	357345 358463 +	T2T	PF10_0084	360625 362480 -	445	1335 H2T	PF10_0085	365737 367146 -	3257	2162	100	100	0	0	0	0	0



Sample cumulative % fail rate file for different UTR lengths

	100	99	98	97	96	95	94	93	92	91	90	89	88	87	86	85	84
600bp UTR	25.245098	25.087535	24.807423	24.6498599	24.4747899	24.2647059	24.1771709	23.9845938	23.9845938	23.9670868	24.0021008	23.7570028	23.7394958	23.5469188	23.4243697	23.4943978	23.5644258
800bp UTR	38.0427171	37.762605	37.67507	37.535014	37.272409	37.1148459	36.9397759	36.5721289	36.5721289	36.8421569	36.3620448	36.1969748	36.2570028	36.1519608	36.2219888	36.2745098	36.0819328
1000bp UTR	51.0854342	51.0154052	50.6827731	50.3851541	49.9649986	49.6673659	49.5448179	49.4047619	49.3347339	49.0721209	49.0021008	48.7394958	48.7044418	48.5459188	48.4943978	48.3718487	48.6954678
1200bp UTR	62.0798319	61.8172269	61.5196078	61.3620448	61.1169468	60.8368347	60.8192277	60.6267507	60.5392157	60.2941178	60.2065928	60.1540616	60.1890756	60.2941176	60.1365546	60.3891597	60.4341737
1400bp UTR	70.6757703	70.3606443	70.2556022	70.0280112	69.9229692	69.7829132	69.6428571	69.3977591	69.1876751	68.9425777	68.9550958	68.9959098	69.012605	68.872549	68.7145986	69.0651261	69.5203081
1600bp UTR	78.1687675	77.9761905	77.8186275	77.4859944	77.1883754	77.0308123	76.8907563	76.7507003	76.4880952	76.2955182	76.1554622	76.0154062	75.9978992	76.2254902	76.4705882	76.6981793	76.9432773
1800bp UTR	82.9481793	82.8431373	82.7906162	82.6680672	82.4579832	82.3879552	82.2478992	82.1603641	82.0378151	81.8452381	81.9152661	82.0203081	82.1078431	82.3529412	82.3529412	82.4229692	82.6330532



2.7.4 POLY

POLY is a bioinformatic tool that calculates parameters for non-overlapping homopolymer tracts in any given sequence. It was written in the scripting language Python by Bizzaro and Marx (2003) and runs in a UNIX environment (Bizzaro and Marx, 2003). Essentially, genomic sequence data (of length x) is input and POLY measures the frequency at which the homopolymeric tract is observed in any given sequence and its relative length giving an output of over-proportionment ($\log(f_{\text{observed}}/f_{\text{predicted}})$) against the polymer length: N for dA, dT, dG and dC tracts. A second variable, R , over-representation, is the log frequency of length, N , of the poly (dA, dT, dG or dC) tract observed over the frequency determined in a random sequence of equivalent nucleotide composition. N_{obs} and N_{exp} represent the maximal size of poly (dA, dT, dG or dC) observed and expected respectively for an equivalent random sequence. P , over-proportionment, is the ratio of $N_{\text{obs}}/N_{\text{exp}}$. POLY is open source software and available at <http://bioinformatics.org/poly/>. In this instance, POLY was used to analyse three different genomic sequence domains; coding sequence (CDS), the 5' upstream region and the 3' downstream region for three different *Plasmodium spp.* of varying genomic AT-content namely; *P. falciparum*, *P. knowlesi*, and *P. vivax*. The workflow and example output data are shown in Figure 5. These analyses were then extended to encompass other *Plasmodium* and Apicomplexan parasites. Two additional scripts were also especially written in order to provide controls for the output.

Figure 2-5 Schematic representing the workflow for use of POLY.

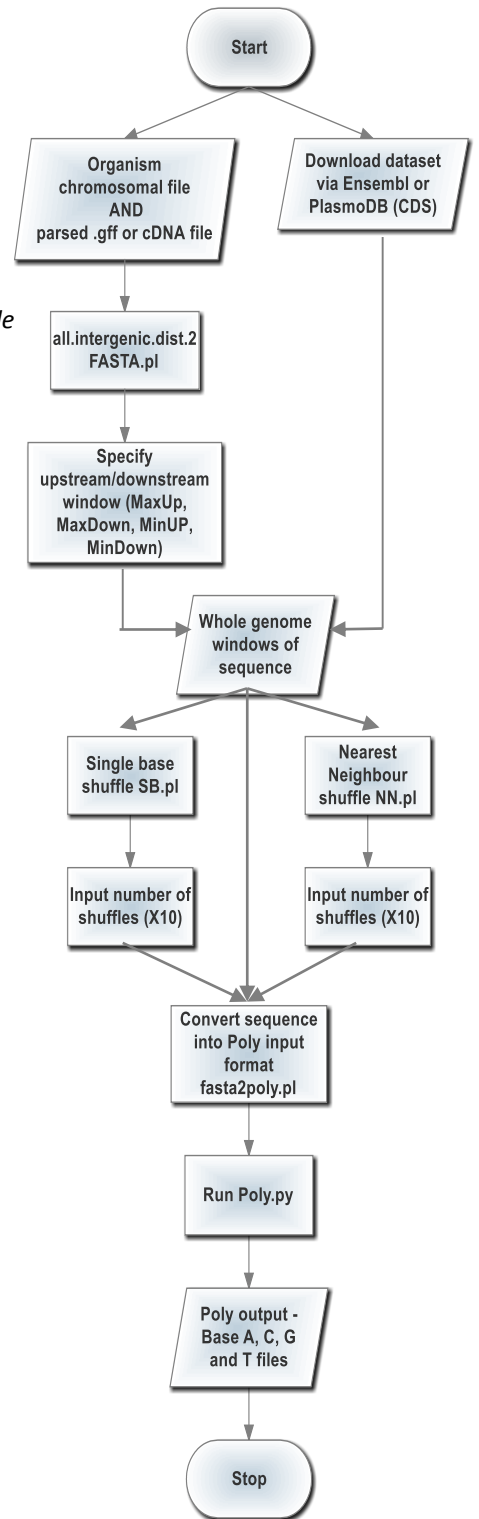
(overpage) Data windows were either extracted from the organism chromosomal files and the parsed .gff (or cDNA) file using the Perl script all.intergenic.dist.2FASTA.pl (written by Dr Richard Emes) or, data was downloaded directly from Ensembl or PlasmoDB. These windows of sequence were converted into a POLY-compatible format with fasta2poly.pl (written by Dr Richard Emes) and run through Poly.py (Bizzaro and Marx, 2003). For *P. falciparum* data, POLY input windows of sequence were also subject to 'shuffling' scripts; a single base shuffle (SB.pl) and a nearest Neighbour shuffle (NN.pl) (written by Chai-Ho Chen) and the POLY analyses also run upon these datasets.

Sample Base A Poly output file

1	203581	-0.74	0.112
2	39742	-1.45	-0.219
3	20660	-1.73	-0.126
4	8915	-2.1	-0.113
5	4139	-2.43	-0.068
6	1760	-2.8	-0.062
7	1009	-3.04	0.074
8	588	-3.28	0.218
9	382	-3.46	0.408
10	280	-3.6	0.651
11	224	-3.7	0.932
12	191	-3.76	1.24
13	153	-3.86	1.522
14	130	-3.93	1.829
15	120	-3.97	2.172
16	125	-3.95	2.567
17	110	-4	2.89
18	81	-4.14	3.135
19	93	-4.08	3.572
20	54	-4.31	3.714
21	50	-4.35	4.058
22	47	-4.37	4.409
23	53	-4.32	4.839
24	54	-4.31	5.225
25	36	-4.49	5.427
26	47	-4.37	5.92
27	43	-4.41	6.26
28	36	-4.49	6.56
29	29	-4.58	6.844
30	28	-4.6	7.207
31	22	-4.7	7.48
32	22	-4.7	7.858
33	26	-4.63	8.308
34	16	-4.84	8.475
35	20	-4.74	8.95
36	12	-4.97	9.106
37	17	-4.82	9.635
38	14	-4.9	9.928
39	8	-5.14	10.06
40	8	-5.14	10.44
41	5	-5.35	10.61
42	2	-5.74	10.59
43	5	-5.35	11.37
48	2	-5.74	12.86
51	2	-5.74	13.99

Sample Base A, T, G and C Poly output file

	A	T	G	C
1	0.11	0.11	-0.05	-0.06
2	-0.18	-0.19	0.17	0.21
3	-0.15	-0.15	0.49	0.53
4	-0.09	-0.11	0.95	1.03
5	-0.04	-0.07	1.52	1.65
6	-0.02	-0.05	2.16	2.32
7	0.1	0.08	2.82	3.09
8	0.23	0.18	3.67	3.82
9	0.42	0.37	4.68	4.95
10	0.65	0.58	5.51	5.62
11	0.85	0.75	6.56	6.51
12	1.14	1.03		7.87
13	1.44	1.32		
14	1.77	1.63		10.25
15	2.09	1.96		
16	2.41	2.23		
17	2.72	2.54		
18	3.07	2.84		
19	3.37	3.16		
20	3.65	3.42		
21	3.99	3.74		
22	4.27	3.99		
23	4.62	4.36		
24	5	4.67		
25	5.33	5.02		
26	5.61	5.29		
27	5.97	5.63		
28	6.33	5.95		
29	6.65	6.25		
30	6.86	6.52		
31	7.18	6.72		
32	7.48	7.02		
33	7.75	7.32		
34	8.07	7.57		
35	8.36	7.88		
36	8.71	8.23		
37	9.08	8.47		
38	9.32	8.84		
39	9.69	9.08		
40	9.76	9.33		
41	9.99	9.44		
42	10.16	9.55		
43	10.48	9.69		
44	10.8	9.62		
45	10.75	10.09		
46		10.15		
47	11.49	10.5		
48				
49	12.06	11.21		
50	12.43	11.56		
51		12.09		
52				
53	13.54			

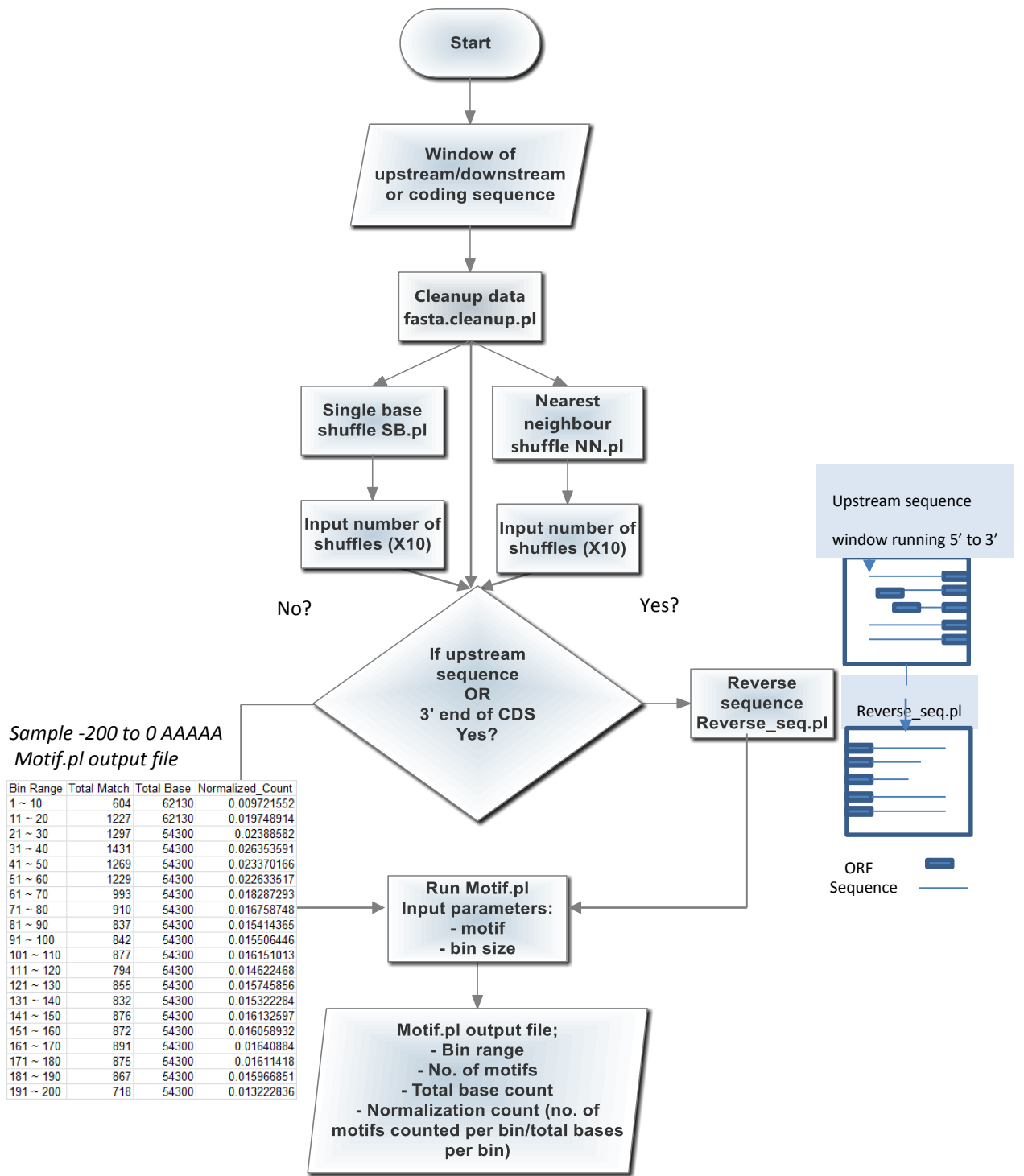


2.7.5 ORGANISATION OF HOMOPOLYMERIC dA.dT TRACTS AROUND THE ORF

Motif.pl was a Perl script written to identify the spatial distribution of user-specified 'motifs' within given 'windows' of sequence (unpublished data – scripts written by our collaborators Chai-Ho Chen and Kenneth Marx of the University of Massachusetts at Lowell, USA). Motif.pl runs in UNIX environment and uses genomic input sequence of length x and allows the user to define their 'motif' of interest and a bin size of their choice. The output file comprises the bin range, the total number of matches, the total base count and a normalized count of the motif incidence. Motif.pl was used, in this instance, to identify the location of homopolymeric (dA) or (dT) tract 'motifs' of 5, 10, 15, 20 or 25bps in length within intergenic flanking regions of *P. falciparum*, *P. vivax* and *P. knowlesi* initially, then subsequently, intergenic regions of other *Plasmodium* species and Apicomplexan parasites. See Figure 6 for workflow and example output data. Two additional scripts were also especially written in order to provide controls for the output. The first script, SB.pl, took the input sequences and subjected them to a single base shuffle (in this case X10). The output file was then run through Motif.pl and used as a comparison/control. The second script, NN.pl, again 'shuffles' the input sequence (X10) but also takes into account the nearest neighbour preferences within the input sequence. The output data from NN.pl, when run through Motif.pl, was used as a secondary comparison/control.

Figure 2-6 Schematic illustrating the workflow for the use of Motif.

(overpage) Windows of upstream or downstream sequence (usually taken from PlasmoDB – specified in Chapter 6 otherwise) were subject to fasta.cleanup.pl (written by Dr Richard Emes) to ensure data compatibility with Motif.pl. For the *P. falciparum* data Motif.pl input windows of sequence were also subject to the shuffling scripts SB.pl and NN.pl (written by Chai-Ho Chen) to provide comparative controls. As not all sequence data within the specified window were full-length, usually owing to an upstream ORF clash, and to ensure that the data analysed flanked the ORF 5' sequences were subject to reverse_seq.pl (see inset figure). The output data was shown as a Normalized_Count – the frequency of total matches (# motifs counted) within a bin divided by the total bases counted in that bin. The normalized count was then plotted against the bin window (bps).



CHAPTER 3 ORGANIZATION OF INTERGENIC SPACE IN *P. FALCIPARUM* AND OTHER APICOMPLEXAN PARASITES

3.1 INTRODUCTION

The current paradigm for transcription in *P. falciparum* invokes a model of monocistronic transcription that initiates and terminates within the flanking intergenic regions (IGR). The promoter is thought to share a canonical bipartite structure with those of other eukaryotes, with promoter deletion studies (reviewed in Horrocks *et al.*, 2009; Wong *et al.*, 2011), suggesting that *cis*-acting regulatory elements, (mainly) located upstream of a transcriptional start site - probably in conjunction with epigenetic mechanisms, translational control and global RNA Polymerase II activity, all contribute to the expression of its gene repertoire (Shock *et al.*, 2007; Sims *et al.*, 2009; Cui and Miao, 2010; Hoeijmakers *et al.*, 2012b). Whilst it is recognised that key transcriptional landmarks for transcript initiation and termination are poorly described in *P. falciparum*, what is not so well recognised is that our wider understanding of the size and orientation of the flanking IGR is similarly limited.

In 2005, Szafranski *et al.*, investigated IGR spacing apportionment for several organisms, including the then incomplete *P. falciparum* chromosome 3, to investigate this phenomenon (Szafranski *et al.*, 2005). Their work suggested a 3:2:1 IGR spacing rule for compact genomes - whereby IGRs containing two promoter regions were larger than those containing a promoter and a terminator region which were in turn larger than an IGR containing two terminator regions. In addition, studies in *Saccharomyces cerevisiae* and other fungal species, as well as *Escherichia coli*, find that upstream control regions (UCR), those containing two promoters, are larger than downstream control regions (DCR) containing two terminators (Hermsen *et al.*, 2008). Together, these studies suggest that IGRs within compact genomes

may be 'spaced' to accommodate the different components of the transcriptional units that reside within them.

This chapter describes a complete characterization of the IGRs for the whole *P. falciparum* genome. In addition, the analyses presented here go beyond that of Szafranski and colleagues to include the complete genomes of *S. cerevisiae* and *D. dictostelium* (included as partial genomes within their study) and also includes analysis of IGR size and organisation for ten other Apicomplexan genomes. Given the distinct function of genes residing in different *P. falciparum* chromosomal compartments (subtelomic v chromosome internal) a comparative analysis of variations in the size of the intergenic space within these distinct compartments is also undertaken. This initial work provides an underpinning for work presented in subsequent chapters of this thesis which explore transcript size and apportionment over these IGRs by defining the spatial organisation of IGRs in terms of the transcriptional activity that takes place over them.

3.2 CAPTURE AND CHARACTERIZATION OF INTERGENIC DATASETS

Annotated *P. falciparum* genomic data are publicly available from PlasmoDB. *P. falciparum* release 5.3, the latest version at the time the work was undertaken, provided a general format file (.gff) (<http://plasmodb.org/plasmo>). This file contained, amongst a lot of other information, the necessary ORF chromosomal co-ordinates and the strand upon which each ORF resides. Using the start and stop co-ordinates for each gene pair, the size of the IGR could be determined. Each IGR could then be categorized into one of four orientation groups based on the orientation of its flanking genes: 5' – 5' or Head to Head (HH), 5' – 3' or Head to Tail (HT), 3' – 5' or Tail to Head (TH) and 3' – 3' or Tail to Tail (TT). The workflow for securing this information is described in Fig. 2-1 (Chapter 2 Materials and Methods) with a more detailed description of this process shown here (Fig. 3-1). This process was automated via the

use of two Perl scripts; `plasmoDB.gff.sorter.pl` and `intergenic.dist.pl`. These two scripts were kindly written by Dr Richard Emes and used by myself to extract the appropriate information from the *P. falciparum* genome and all subsequent genomes analysed.

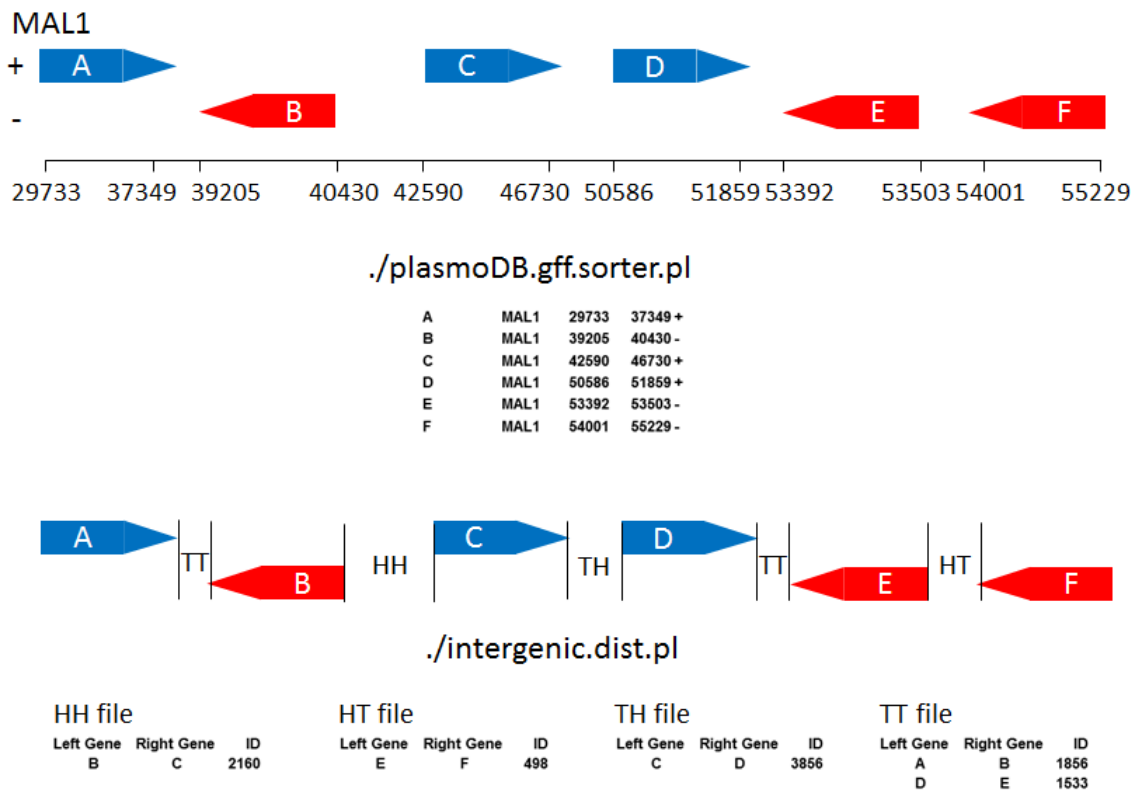


Figure 3-1 Overview of the extraction of relevant information from a .gff file and allocation of the intergenic distance to the relevant gene orientation file

Intergenic distance co-ordinates were secured from the PlasmoDB v5.3 general format (.gff) file. Using the Perl script `plasmoDB.gff.sorter.pl`, relevant information is extracted from the .gff file to give a parsed file containing the gene identifier (e.g. for genes A to F), the chromosome number (all on chromosome 1, MAL1), the start and end co-ordinates of the ORF and the DNA strand on which it resides (where + = Watson strand and - = Crick strand). The Perl script `Intergenic.dist.pl` was then used to extract the intergenic distance (in bp) between pairs of adjacent ORF on each chromosome sequentially. The corresponding intergenic distances between the pair of ORFs were then allocated to the appropriate orientation file (HH, HT, TH or TT) using the DNA strand information.

3.3 ANALYSIS OF INTERGENIC SPACE IN *P. FALCIPARUM*

The numbers of each IGR type and the sizes of 5588 *P. falciparum* intergenic distances were captured via this methodology. Of these, 1479 were flanked by genes in the 5'–5' or HH orientation, 1274 were flanked by genes in the 5'–3' or HT orientation, 1352 were flanked by genes in the 3'–5' or TH orientation and 1483 were flanked by genes in the 3'–3' or TT orientation. These data conformed, more or less, to the anticipated 1:1:1:1 group ratio expected for monocistronic transcription (Lanzer *et al.*, 1992a). The range of the values for each subgroup (HH, HT, TH and TT) is shown in Table 3-1. It should be noted that a small number of excessively high and low values were present within each dataset and that none of the datasets were normally distributed (Fig. 3-2).

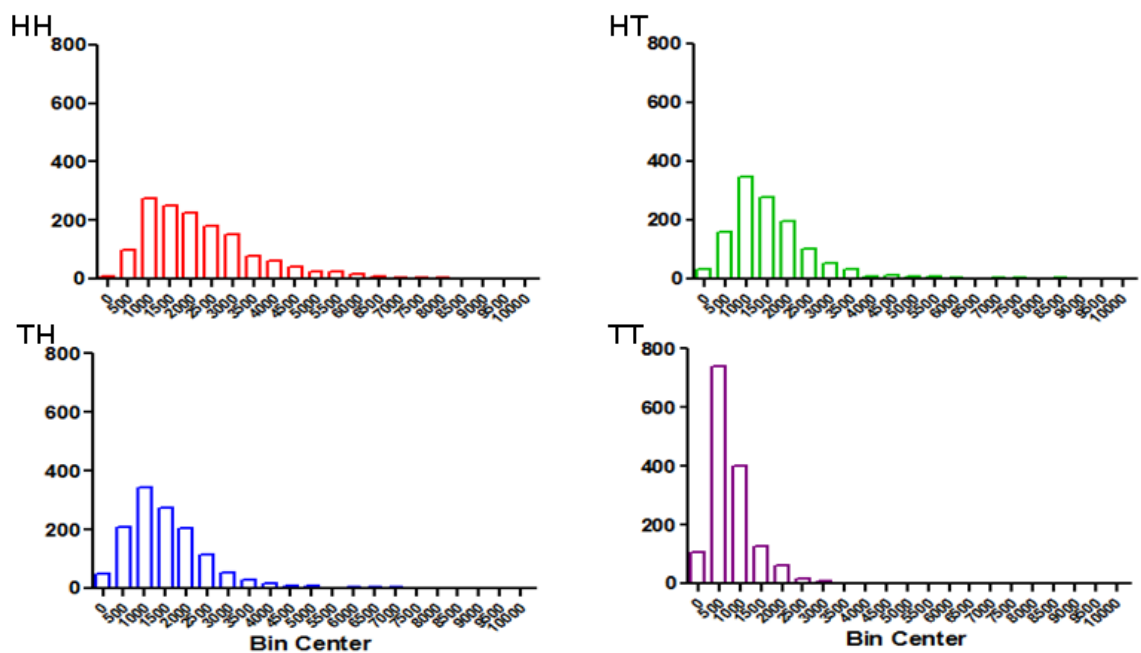


Figure 3-2 Frequency histogram representation of *P. falciparum* intergenic distance data when sub-grouped by flanking gene orientation

HH – intergenic distance flanked by two genes in the 5' orientation, HT and TH – intergenic sequence flanked by one gene in the 5' orientation and one gene in the 3' orientation, TT – intergenic sequence flanked by two genes in the 3' orientation. The histograms (bin size of 500bp) for the *P. falciparum* IGR orientation groups demonstrate an inherent skew within and non-normal distribution for each of the datasets.

Table 3-1 Range of values within each intergenic distance dataset

<i>Data Range (bps)</i>							
HH		HT		TH		TT	
Min	Max	Min	Max	Min	Max	Min	Max
20	22869	2	11087	1	21048	9	14335

As none of the datasets demonstrated a normal distribution (Shapiro Wilk Test) – all had an inherent skew, this meant that either all datasets would have to be transformed or non-parametric analyses utilized for the comparative analyses. As there was no biological imperative that the data should be of a normal distribution and no one transformation successfully converted all datasets, non-parametric analysis was employed. Table 3-2 shows the difference between the mean and the (non-parametric) median for each dataset.

Table 3-2 Mean and Median values for each intergenic distance dataset

<i>Flanking gene orientation</i>	<i>Mean (bps)</i>	<i>Median (bps)</i>
HH	2305	1938
HT	1684	1407
TH	1623	1366
TT	854	677

A non-parametric ANOVA (Kruskal-Wallis with a Dunns Multiple Comparison Post Test) was used to evaluate whether the groups were equivalent. The Dunns Multiple Comparison Post-test established that there was a statistically significant difference ($p < 0.05$) between all groups with the exception of the HT and TH groups. This was not surprising as HT or 5'-3' and TH or 3'-5' IGRs can be regarded as functionally equivalent i.e. they both contain one promoter and one terminator. Fig. 3-3 represents these data as a Box and Whiskers plot. Whilst there is a clear difference in the size of the IGR depending upon the orientation of the flanking genes, with a $HH > HT = TH > TT$ relationship, these differences are clearly obscured by the presence of a few 'outliers'. These outliers likely represent artefacts arising from errors in gene annotation or IGR that encompass regions such as centromeres. How these differences

can be more effectively illustrated is discussed below after a rationalisation of the datasets into three functionally-relevant groups.

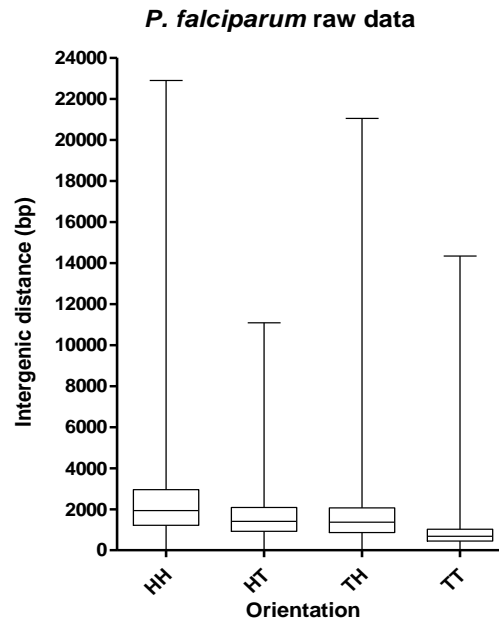


Figure 3-3 *P. falciparum* intergenic data grouped upon flanking gene orientation

Box and whiskers plot of the four IGR datasets (HH, HT, TH and TT) which demonstrates a clear $HH < HT = TH < TT$ pattern of median IGR distance according to the orientation of the flanking genes. The plot also shows the min and max data values.

As no significant difference was observed between the HT and TH datasets (Mann Whitney U test, $P = 0.0976$) and these two datasets represent IGR with equivalent transcriptional activity occurring over them (i.e. one promoter and one terminator), the HT and TH datasets were concatenated and the datasets described as IGR types A - (HH), B - (HT and TH) and C (TT), where transcripts for the flanking ORF over these IGR are either divergent, tandem or convergent, respectively (see Fig. 3-4 for schematic). These datasets are depicted graphically as a frequency histogram in Fig 3-5A and as a box plot in Fig 3-B.

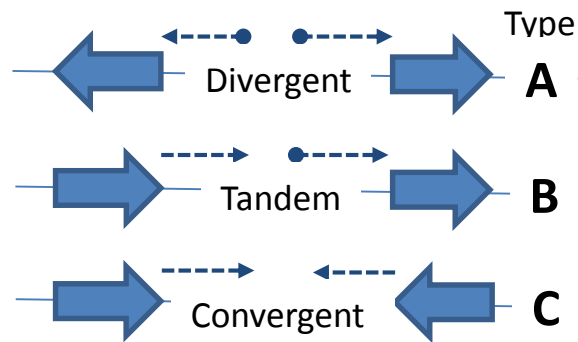


Figure 3-4 File concatenation and re-termining of datasets

The four IGR datasets were amalgamated into three; A Divergent (containing two promoters), B Tandem (containing one promoter and one terminator) and C convergent (containing two terminator) according to the transcriptional units which reside within them.

As it was also observed that a small proportion of the data for each IGR type were of an excessively high or low value (Fig. 3-3) graphical depictions of 98% (2-99% values, data not shown) or 95% (2.5-97.5% values, Figure 3-5C) were considered. Although both the 98% or the 95% datasets would have been suitable for the *P. falciparum* dataset, subsequent analysis of other Apicomplexan genomes, many of which had far less complete genomic datasets (some still in contig format), indicated that a 5% reduction (i.e. display 2.5-97.5%) in data was more appropriate for comparative purposes. It should be noted that all analyses carried out here were of a non-parametric nature as none of the datasets were normally distributed. Therefore, all measurements were about the median, not the mean, thus the removal of 2.5% of data from either end of the dataset would not affect any subsequent results (see Table 3-3).

Table 3-3 Rationalisation of intergenic datasets

Total data (HT/TH combined)					
100% data					
	Number	Median	Mean	SD	SE
A	1479	1938	2305	1577	41.00
B	2626	1385	1652	1266	24.71
C	1483	677	854	844	21.91
95% data					
	Number	Median	Mean	SD	SE
A	1405	1938	2208	1200	32.02
B	2494	1385	1557	867	17.36
C	1409	677	781	437	11.65

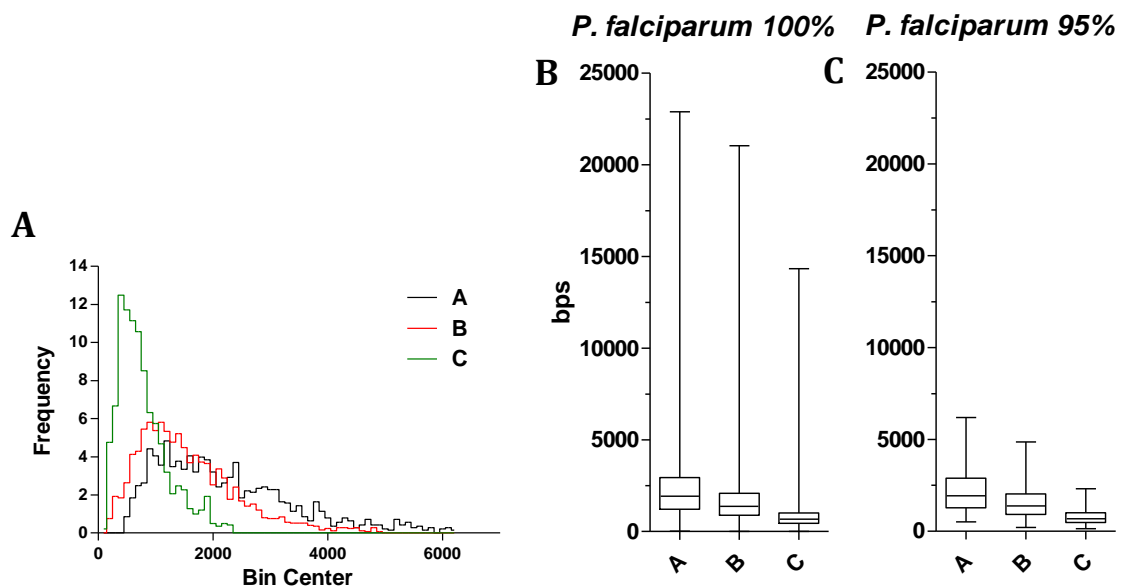


Figure 3-5 Frequency histogram of the *P. falciparum* IGR types A,B,C and boxplots showing 100% and 2.5-97.5% of the data respectively

(A) Frequency histogram (bin size 100bp) demonstrates the distribution of IGR size arrangement types; A, B and C (see key). (B) and (C) Box and whisker plots of the size of IGR within types A, B and C demonstrating the effect of removing extreme values from either ends of the dataset; the 95% dataset, (showing 2.5-97.5% of data), effectively removing outliers.

A 1:1.8:1 A:B:C relationship for the IGR count within each category is demonstrated (data from Table 3-3) and this again is close to the 1:2:1 ratio that would be predicted for independent monocistronic transcription units. The median A:B:C size ratio, with type C IGR

taken as 1, is 2.86:2.05:1 (Table 3-3) therefore, the size of a type A IGR > type B IGR > type C IGR in an approximate to 3:2:1 ratio. These analyses confirmed that for *P. falciparum*, on a genome wide scale, there was a clear difference between the median length of the IGR according to the transcriptional units it contained. These data strongly suggest that within the compact *P. falciparum* genome more intergenic space is required to accommodate promoter regions than terminator regions.

3.4 INTERGENIC DISTANCES ARE LONGER IN *P. FALCIPARUM* SUBTELOMERIC REGIONS

P. falciparum chromosomes can be broadly considered to contain three functionally distinct compartments: (i) Subtelomeric – which tend to have a reduced gene density, contain arrays of multi-gene families often associated with virulence (such as the *var*, *rif* and *stevor* gene families) and are typically organized within heterochromatin (Gardner *et al.*, 2002; Hoeijmakers *et al.*, 2012b). (ii) Chromosome internal - which contains the majority of ORFs in the genome. Many of these genes are single copy, share some degree of synteny with genes from chromosomal internal regions of other *Plasmodium spp.* and are predominantly organized within euchromatin (Gardner *et al.*, 2002; Ponts *et al.*, 2010). (iii) Centromeres – a single region of each chromosome of approximately 3kbp, rich in repetitive AT-rich sequences, containing no ORFs (Hoeijmakers *et al.*, 2012a). This compartment is not considered further within these analyses.

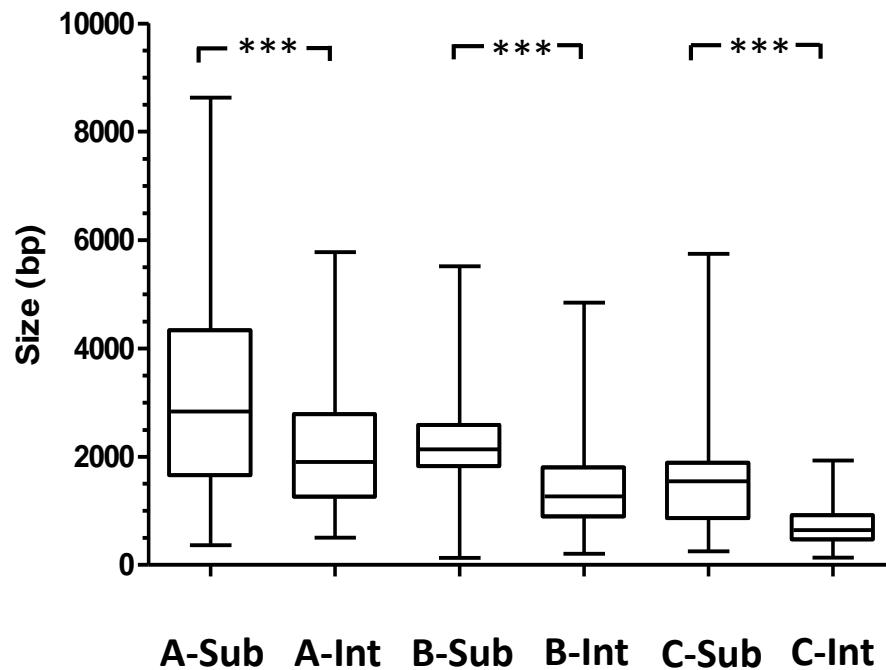
A significant amount of research has concentrated on subtelomeric regions in *P. falciparum* owing to the fact that the majority of this parasite's *var* genes, which encode one of the major virulence proteins, PfEMP1, are located here. Through a process of mutually-exclusive expression, predominantly mediated via epigenetic means, a single *var* gene variant is expressed (Freitas-Junior *et al.*, 2005; Dzikowski *et al.*, 2006; Scherf *et al.*, 2008; Petter *et al.*,

2011). Periodic 'switching' to another *var* variant results in a change of the PfEMP1 protein expressed on the surface of the iRBC which delays immune recognition (Pasternak and Dzikowski, 2009). Recombination of subtelomeric virulence genes plays an important role in the survival of this parasite, generating antigenetically-diverse repertoires to be called upon in future generations (Templeton, 2009). Therefore, subtelomeric regions tend to accommodate genes with different functional requirements and may be considered to be subject to very different pressures than most of the *P. falciparum* gene repertoire. The telomeric regions, as well as the non-coding regions immediately adjacent to the telomeres, are also, as in most other organisms, subject to length plasticity (Figueiredo *et al.*, 2002; Hernandez-Rivas *et al.*, 2010).

To explore the relationship between intergenic distances within subtelomeric and chromosomal internal compartments, IGR breakpoints were established between chromosome internal genes - deemed as those genes for which synteny with *P. vivax* and *P. knowlesi* existed and the species-specific genes located within subtelomeric regions (see Table 3-4 for the chromosomal breakpoints and Fig. 2-2 (Materials and Methods) for an overview of the process). The IGRs for each chromosomal compartment were then split into the three IGR orientation groups; A, B and C respectively to give six data groups for the comparative analysis (see results in Fig. 3-6).

Table 3-4 Chromosomal breakpoints deemed to be chromosome internal and subtelomeric.

<i>Chromosome</i>	<i>Subtelomeric left</i>	<i>Chromosome internal left</i>	<i>Chromosome internal right</i>	<i>Subtelomeric right</i>
1	PFA0135w	PFA0140c	MAL1_28S:rRNA	PFA0610c
2	PFB0115w	PFB0120w	PFB0905c	PFB0910w
3	PFC0090w	PFC0095c	PFC1065w	PFC1070c
4	PF01144c	PF01145c	PF01115c	PF01120c
5	PFE0070w	PFE0075c	MAL5_28S:rRNA	PFE1590w
6	PF0090w	PF0095c	PF01505w	PF01510w
7	PF07_0009	MAL7_tRNA_Ala1:tRNA	MAL7_tRNA_Val1:tRNA	MAL7P1.170
8	MAL8P1.160	PF08_0137	MAL8P1.4	MAL8P1.3
9	PFI0140w	PFI0145w	PFI1730w	PFI1735c
10	PF10_0026	PF10_0027	PF10_0420	PF10_0374
11	PF11_0042	PF11_0043	MAL11_rRNA:rRNA	PF11_0503
12	PFL0070c	PFL0075w	PFL2505c	PFL2510w
13	MAL13P1.62	PF13_0076	PF13_0361	PF13_0362
14	PF14_0020	PF14_0021	PF14_0725	PF14_0726



Region	IGR Type	n=	Ratio of IGR		% change	Ratio of median size
			types	Median Size		
Subtelomeric	A	123	0.98	2838	+46.4	1.84
	B	379	3.03	2138	+54.4	1.38
	C	125	1.00	1545	+128.2	1.00
Internal	A	1283	1.01	1905	-1.1	2.95
	B	2118	1.66	1266	-8.6	1.96
	C	1276	1.00	646	-4.6	1.00

Figure 3-6 Comparison of the size of chromosomal internal IGRs with subtelomeric IGRs.

Figure 6 demonstrates the increase in the median A, B and C IGR size for the subtelomeric datasets when compared to chromosome internal A, B and C IGR datasets with the increase in size of the subtelomeric IGRs rising to 128% for the C dataset. The 3:2:1 A:B:C ratio previously observed for the whole genome also reduces to a 1.8:1.4:1 ratio within the subtelomeric dataset.

11.8% or 627 of the total 5588 genomic IGRs were deemed here to be subtelomeric. This gave a subtelomeric A:B:C count of 123:379:125 and a A:B:C ratio of 1:3:1 - in contrast to ~1:2:1 for the chromosome internal compartment. The large increase in the number of B type IGRs observed is probably attributable to a head to tail bias for the numerous and subtelomerically

located *rifin* family of genes (Kyes *et al.*, 1999; Joannin *et al.*, 2008). The 3:2:1 A:B:C IGR spacing ratio observed in the genomic dataset also collapses to a 1.8:1.4:1 ratio for subtelomeric IGRs – principally as the type C IGR has the greatest increase in length. This reduction in the ratio suggests that the distinct function of subtelomerically located gene families has a significant impact on IGR space requirements – essentially a less dense gene organization with more IGR available.

3.5 THE LEVEL OF TRANSCRIPTIONAL ACTIVITY DOES NOT APPEAR TO CORRELATE WITH THE SIZE OF THE FLANKING IGR

Otto *et al.*, (2010) recently published RNASeq data for 4871 *P. falciparum* genes transcribed during the IE cycle. These data provide a wealth of information, amongst which is the mRNA abundance (mRNAa) - a quantitative assessment of steady-state mRNA from 7 time-points over the IE cycle. Determining the geometric mean of these 7 time-points for each gene gives an average level of mRNAa for each gene over this entire cycle, thus approximating the overall mRNA abundance (Otto *et al.*, 2010). These data enabled the comparison of the mean gene transcription activity with the size of the flanking type A, B and C IGR.

Prior to allocating the mean mRNA abundance level to each gene flanking region, those genes with ORFs containing large low complexity regions were excluded from the analysis. Otto *et al.* (2010) reports an inability to unambiguously allocate short RNASeq reads to genes containing these regions and as such this could introduce bias (i.e. these genes would appear relatively under-abundant) (Otto *et al.*, 2010). This resulted in 1093 of 4871 genes for which RNASeq data from the IE cycle were available being excluded. As each gene has two flanking sequences, the mRNA abundance data for each gene were used twice in this analysis. Note due to the distribution of mean abundance and intergenic distances scatterplot correlations use a Log₁₀ scale (Fig. 3-7 A, B and C). These data were also plotted as box and whisker plots

using quartiles of the intergenic distance to facilitate an ANOVA analysis for significant differences between each group (Fig. 3-7 D, E and F).

No linear correlation between the sizes of intergenic types A, B or C and mRNA abundance was found (all $R^2 < 0.05$). The scatterplot analyses did seem to suggest a trend where the most abundant mRNA were derived from longer intergenic regions. This is not likely to be simply due to larger regions being available for RNA-Seq reads to be annotated to, as these reads were rarely annotated outside of the open reading frame sequences. The quartile analysis does suggest that mRNA abundance in the largest quartile of intergenic distance length is always significantly higher ($P < 0.05$ ANOVA with Dunns post-test) although from Table 3-5 it is clear that this is actually only a small difference.

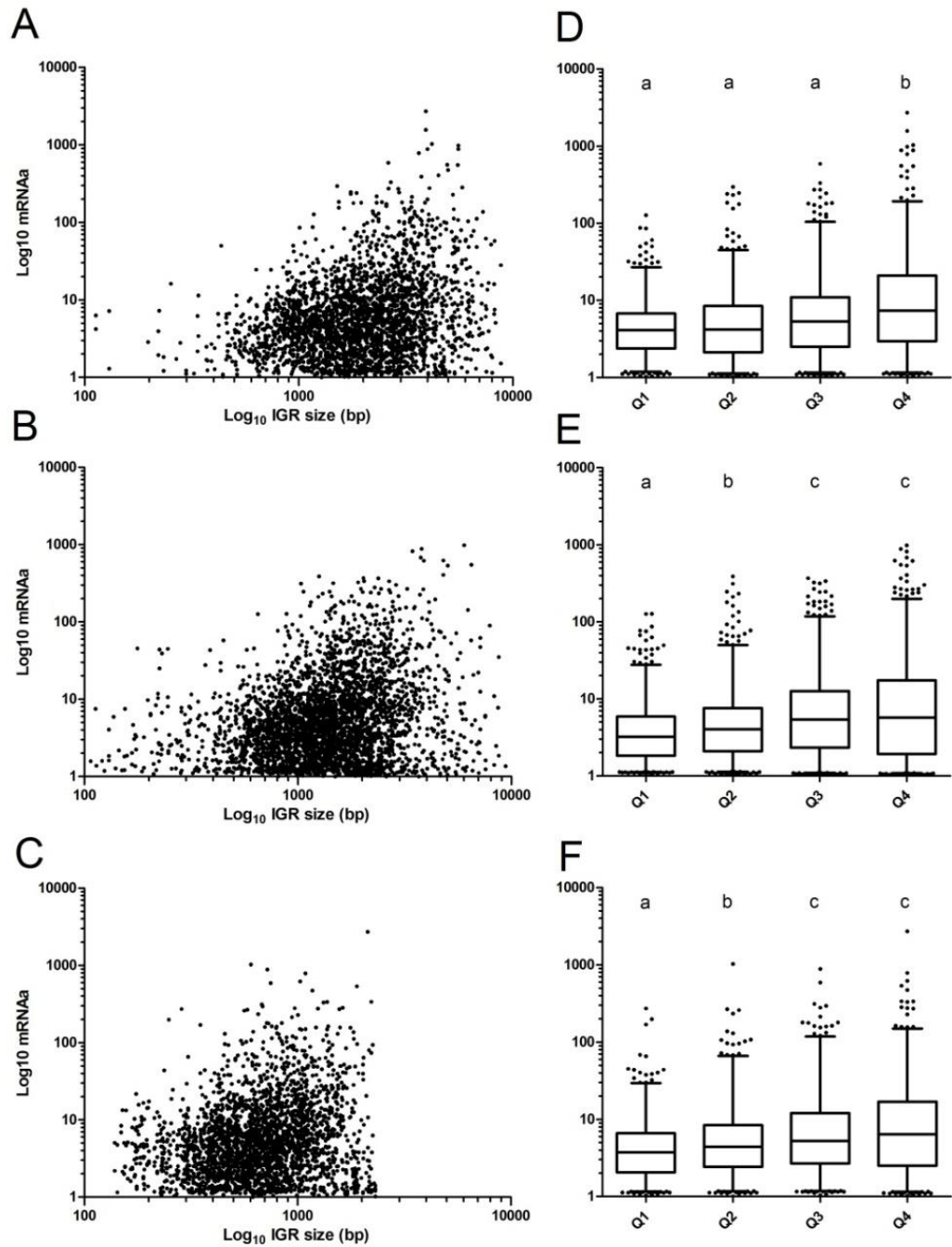


Figure 3-7 Correlation between the size of the IGR and mRNA abundance (mRNAa)

The scatterplots show the Log_{10} transformed mRNAa plotted against IGR size (bps). Scatterplots A, B and C represent A, B and C IGR orientation groups respectively. The box and whiskers plots D, E and F show the IGR length split into four quartiles (Q1-Q4) for each IGR orientation group (A, B and C respectively) plotted against mRNAa for IGR within that quartile. ORFs with low complexity regions have been removed from all datasets. The a, b and c values above each box and whisker plot denote whether a significant difference was observed between the quartile groups (ANOVA with Dunn's Post Test, $p < 0.05$).

Table 3-5 Quartile values of IGR length and mRNAa

		IGR length (bp)		mRNAa		
		Min.	Max.	Min.	Median	Max.
IGR A	Q1	113	1236	1.02	4.11	126.7
	Q2	1236	1953	1.01	4.15	294.1
	Q3	1953	2968	1.04	5.28	568.5
	Q4	2968	22896	1.03	7.32	2696
IGR B	Q1	74	921	1.01	3.21	126.9
	Q2	921	1386	1.02	4.03	378.9
	Q3	1386	2046	1.00	5.38	376.0
	Q4	2046	21048	1.00	5.68	279.8
IGR C	Q1	139	455	1.03	3.75	271.2
	Q2	455	692	1.00	4.39	1023
	Q3	692	983	1.04	5.27	881.3
	Q4	983	2316	1.02	6.38	2696

3.6 DIFFERENTIAL APPORTIONMENT OF THE INTERGENIC SPACE IS A FEATURE OF GENOME ORGANISATION IN A RANGE OF APICOMPLEXAN PARASITES

Analyses of *P. falciparum* intergenic datasets demonstrated an A>B>C relationship with a 2.9:2:1 A:B:C median size ratio for genomic intergenic distance. These data suggested that, for this compact genome, the size of the IGR appeared to correlate with the nature of the transcriptional activity that occurs over it. In 2005, Szafranski *et al.*, analysed small datasets from *Dictostelium discoideum*, *Arabidopsis thaliana* and *S. cerevisiae* as well as *P. falciparum*. This group proposed a general IGR 'spacing' rule of 3:2:1 (A:B:C ratio) for genomes with gene densities in the range of 2.5 to 4.8 kb/gene (Szafranski *et al.*, 2005). They also concluded that this rule was independent of genomic nucleotide bias and, interestingly, that within the highly gene dense genome of *S. cerevisiae* that this rule broke down (median ratio 2.0:1.6:1) with little correlation observed between the size of the IGR and the flanking gene orientation (Szafranski *et al.*, 2005).

Using the same approach as described above, the numbers and sizes of all intergenic distances were secured for twelve additional organisms (including *S. cerevisiae* and *D. discoideum* and ten other Apicomplexan parasites). Organism names and information relating to the source of the genomic data are shown in Table 3-6. The raw output data are also presented in Table 3-6 including; the intergenic region (IGR) data count and the data range for each group (HH, HT, TH, TT) for each organism. All datasets were subject to a normality test (Shapiro Wilk) and as for *P. falciparum* - none of these datasets were of a normal distribution (data not shown). Again this meant that non-parametric tests, measuring around the median rather than the mean, had to be employed for all subsequent statistical analysis.

Table 3-6 Raw data download information for the thirteen organisms investigated

Organism	Data source	IGR Count				Data Range (bps)							
		HH	HT	TH	TT	HH		HT		TH		TT	
						Min	Max	Min	Max	Min	Max	Min	Max
<i>B. bovis</i>	Genbank processed A. Pain 23/07/09	1184	1019	1075	1086	12	7136	0	5392	2	7236	0	6329
<i>C. hominis</i>	release-3.7 14/03/08	680	699	649	835	21	3766	1	4494	0	3704	1	4173
<i>C. parvum</i>	release-3.7 20/01/08	1167	912	1006	1142	3	5212	0	7433	0	3954	1	9995
<i>D. discoideum</i>	25/09/08 download	3377	3407	3352	3382	12	10890	3	12298	0	9557	2	8471
<i>N. caninum</i>	release-5.1 25/03/09	1784	1039	1030	1785	152	1350334	2	1745254	4	1461376	180	1316865
<i>P. falciparum</i>	release -5.3 01/05/08	1440	1244	1315	1443	20	22896	2	13532	5	21048	9	18541
<i>P. knowlesi</i>	release-5.5 18/09/08	552	426	460	552	6	13304	2	17708	4	19175	8	33980
<i>P. vivax</i>	release-5.5 19/09/08	1386	1178	1156	1323	6	24363	2	17252	4	15736	0	23027
<i>P. yoelli</i>	Release 5.4 02/06/08	718	1056	1754	1400	3	7673	2	5025	2	8872	3	5139
<i>S. cerevisiae</i>	31/05/08 download	1666	1598	1726	1723	1	7782	0	7380	0	5013	0	5621
<i>T. annulata</i>	Genbank processed A. Pain 23/07/09	915	986	968	903	11	8221	2	2143	3	5282	1	8736
<i>T. gondii</i>	release-4.3/TgondiiME49	2107	1681	1722	2077	2	59755	0	76249	4	60248	1	64839
<i>T. parva</i>	Genbank processed A. Pain 23/07/09	932	1059	1101	908	1	3260	7	6273	1	3227	0	4139

The table shows the release number of the gff file and/or the date of the data download, the quantity of data in each group (HH, HT, TH, TT) and the data range (minimum and maximum) for each group (HH, HT, TH, TT) for each of the ten organisms investigated. HH – intergenic distance flanked by two genes in the 5' orientation, HT – intergenic sequence flanked by one gene in the 5' orientation and one gene in the 3' orientation, TH – intergenic sequence flanked by one gene in the 3' orientation and one gene in the 5' orientation, TT – intergenic sequence flanked by two genes in the 3' orientation. Data was downloaded from <http://dictybase.org/> (*D. discoideum*), <http://plasmodb.org/plasmo/> (*P. falciparum* (3D7), *P. knowlesi* (H), *P. vivax* (Salvador 1), *P. yoelli* (17XNL)), <http://toxodb.org/toxo/> (*T. gondii* (ME49), *N. caninum* (Liverpool)), <http://cryptodb.org/cryptodb/> (*C. hominis* (TU502), *C. parvum* (IOWA)), <http://www.yeastgenome.org/> (*S. cerevisiae*) and Genbank (*B. bovis* (Texas T2Bo), *T. parva* (Mugugu), *T. annulata* (Ankara clone 9)).

Statistical comparison of the intergenic sub-groups (HH, HT, TH and TT) to ascertain whether the median intergenic sizes for each group from each organism were equivalent were carried out using a Kruskal Wallis ($P < 0.05$) with a Dunn multiple comparison post-test and the results are presented in Table 3-7. All comparisons which were considered to be statistically significantly different are annotated Yes, whereas all comparisons which were considered not to be statistically significantly different are annotated No. Analyses that did not follow the expected outcome are highlighted in red text and discussed below. Generally the $HH > HT = TH > TT$ pattern established for *P. falciparum* was repeated as only 5 of the 78 conjectures gave results different from the anticipated pattern.

Table 3-7 Comparison of HH, HT, TH and TT intergenic group medians for the thirteen organisms investigated

	<i>HH v HT</i>	<i>HH v TH</i>	<i>HH v TT</i>	<i>HT v TH</i>	<i>HT v TT</i>	<i>TH v TT</i>
<i>B. bovis</i>	Yes	Yes	Yes	No	Yes	Yes
<i>C. hominis</i>	Yes	Yes	Yes	No	Yes	Yes
<i>C. parvum</i>	Yes	Yes	Yes	No	Yes	Yes
<i>D. discoideum</i>	Yes	Yes	Yes	No	Yes	Yes
<i>N. caninum</i>	Yes	Yes	Yes	No	Yes	Yes
<i>P. falciparum</i>	Yes	Yes	Yes	No	Yes	Yes
<i>P. knowlesi</i>	Yes	Yes	Yes	No	Yes	Yes
<i>P. vivax</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>P. yoelii</i>	Yes	Yes	Yes	No	No	No
<i>S. cerevisiae</i>	Yes	Yes	Yes	No	Yes	Yes
<i>T. annulata</i>	Yes	Yes	Yes	No	Yes	Yes
<i>T. gondii</i>	No	No	Yes	No	Yes	Yes
<i>T. parva</i>	Yes	Yes	Yes	No	Yes	Yes
Statistically significant difference $P < 0.05$ (Kruskal Wallis with a Dunns Post Test)						

The exceptions were; *T. gondii*, (no significant difference between HH and HT or HH v TH), *P. yoelii* (no significant difference between HT v TT or TH v TT) and *P. vivax* (significant difference between HT v TH). For *T. gondii*, this may well be a product of less space constraint as this organism has a larger genome and lower gene density. It should also be noted that although a significant difference was identified for *N. caninum* (which also has a larger genome and lower gene density) this arose owing to the medians of the HT and TH datasets being larger (rather than smaller) than the median value for the HH dataset. Of note for *P.*

yoelli is the difference in distribution of types A, B and C intergenic regions (Fig. 3-8). Presumably this reflects the incomplete nature of the assembly at the time of analysis (still in contig format), also, there appears to be a 50% underrepresentation of type A IGR (see Table 3-6) with the A:B:C numbers being 718:2810:1400 respectively for this organism. Given that 12 of the 13 HT V TH datasets were not significantly different these were all combined to provide a Type B intergenic distance dataset for comparative analysis. Distribution plots for all these additional organisms are shown in Fig. 3-8 and illustrate the anticipated skew in the distribution of intergenic distance size.

When the HT/TH datasets were combined and the A, B and C IGR datasets compiled into box and whiskers plots (Fig. 3-9) a statistically significant difference was observed between the medians of all sub-groups with the exception of *P. yoelii* (B V C) and *T. gondii* and *N. caninum* (A v B). These data confirmed that the strong general trend of $A > B > C$, previously identified in *P. falciparum* is also generally true for other organisms with compact genomes. The two notable exceptions were *T. gondii* and *N. caninum* which do not conform to the $A > B > C$ relationship but instead have an $A = B > C$ relationship. It is worthy of note that both of these coccidian organisms have much larger genomes (64 and 59Mb respectively) and lower gene density (7.4 and 8.1kb/ORF respectively) than the other organisms investigated here. Therefore, for compact genomes (2.0-4.8kb/ORF), the size of the IGR does not appear to be random. Instead, as for *P. falciparum*, the size of the IGR appears dependent upon the nature of the transcriptional activity that occurs over it.

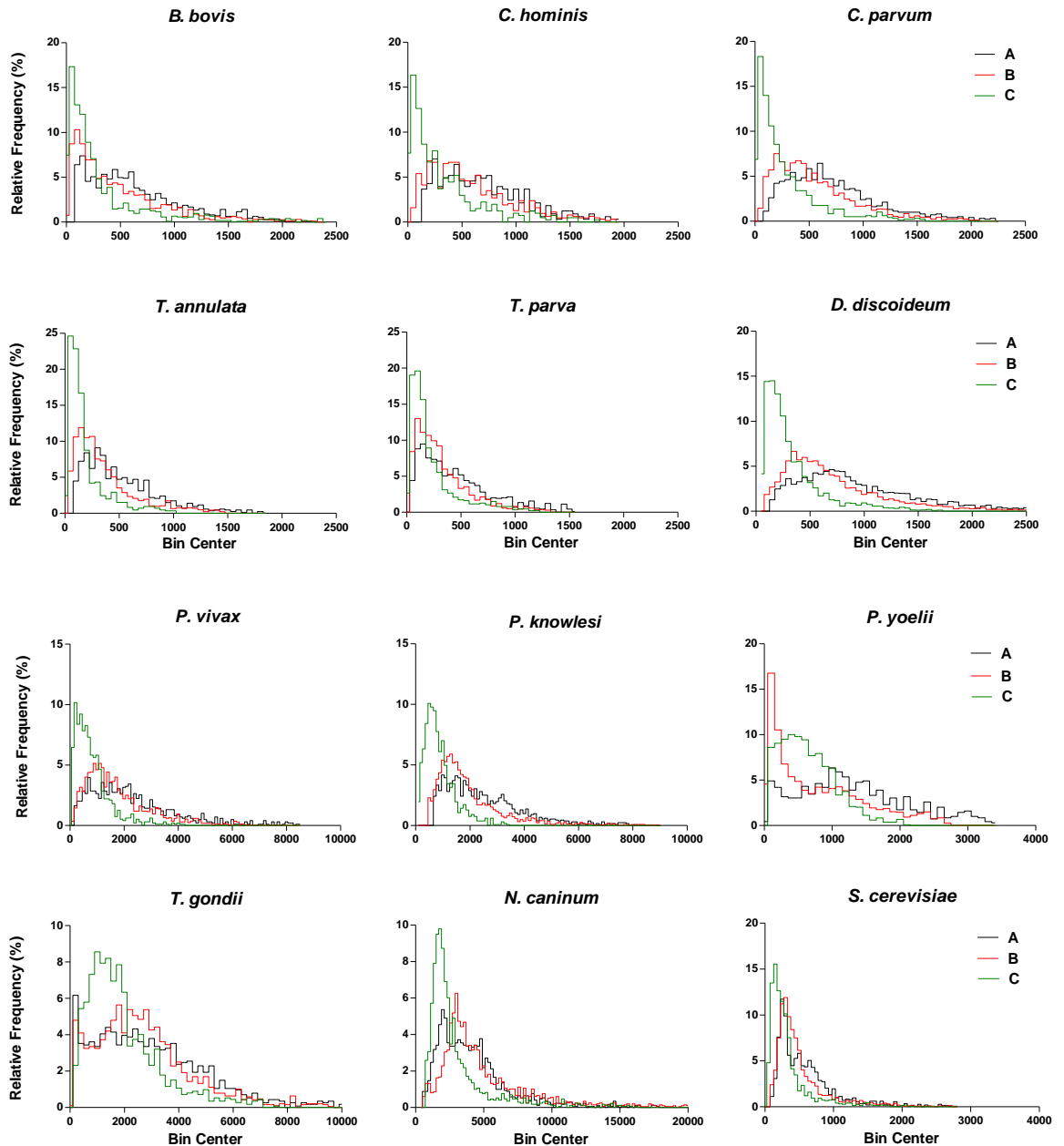


Figure 3-8 Histogram representation of the A, B and C datasets for all organisms

Key: A = Black line, B = red line, C = green line. A = intergenic distances flanked by genes in the HH (5'-5') orientation, B = intergenic distances flanked by genes in the HT (5'-3') or TH (3'-5') orientation and C = intergenic distances flanked by genes in the TT (3'-3') orientation. The histogram representation (GraphPad Prism) demonstrates the data spread over the intergenic distance groups for *B. bovis*, *C. hominis*, *C. parvum*, *T. annulata*, *T. parva*, *D. discoideum*, *P. knowlesi*, *P. vivax*, *P. yoelii*, *T. gondii*, *N. caninum* and *S. cerevisiae*. It should be noted that scales and bin sizes vary. It should also be noted that the genomic data for some of these organisms is incomplete. 2.5 to 97.5% datasets are used for graphical representation.

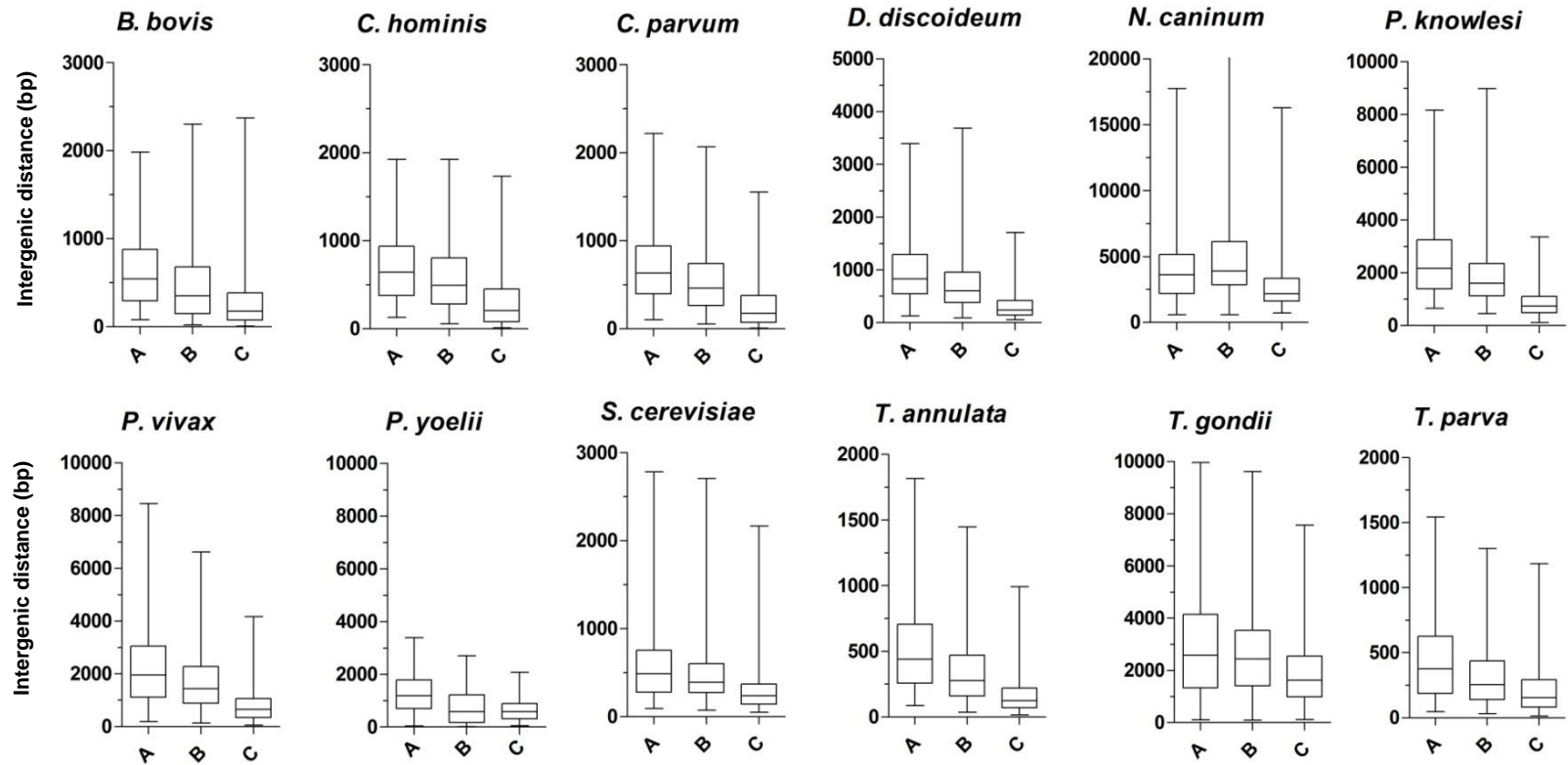


Figure 3-9 Box And Whiskers Plots For The 95% Intergenic Distance Orientation Groups, A, B And C For All Organisms Investigated

(previous page) Box and whiskers plots demonstrating the median intergenic distance and spread of the data for orientation sub-groups A = intergenic distances flanked by genes in the HH (5'-5') orientation, B = intergenic distances flanked by genes in the HT (5'-3') or the (3'-5') orientation and C = intergenic distances flanked by genes in the TT (3'-3') orientation for all ten organisms investigated. The data presented (graphpad prism 5.01) is the rationalised 95% datasets for: *B. bovis*, *C. hominis*, *C. parum*, *T. annulata*, *T. parva*, *D. discoideum*, *P. knowlesi*, *P. vivax*, *P. yoelii*, *T. gondii*, *N. caninum* and *S. cerevisiae*. Note the scales differ and that 2.5 to 97.5% datasets are used for graphical representation.

3.6.1 ANALYSIS OF THE RATIO OF INTERGENIC SPACE ORGANIZATION

Table 3-8 reports the count and median size of each IGR type, and their A:B:C ratio, for each of the organisms investigated here. Monocistronic transcription would dictate an approximate 1:2:1 A:B:C count ratio and indeed most organisms essentially comply with this ratio. *N. caninum* and *P. yoelii* are the major exceptions to this rule, instead showing a 1.0:1.2:1.0 or a 0.5:2.0:1.0 A:B:C IGR count ratio respectively (see Table 3-8). However, it is worthy of note that the data used for the analyses of both of these two organisms were probably the least 'complete', suggesting that this finding may simply be an artefact created by using data produced from contig format rather than compiled genomic sequence.

The ratio of median IGR size is perhaps of more interest. With the exception of those of the coccidian parasites, *N. caninum* and *T. gondii*, an approximate 3:2:1 A:B:C median ratio seems to emerge for all the Apicomplexa analysed here - despite varying genome density, nucleotide content bias and median intergenic A, B and C sizes. For example, whilst for *P. falciparum* this spacing rule is true for a 1939:1385:677 A:B:C IGR ratio, in the much more compact genomes of *B. bovis* and *T. annula* this same spacing rule also seems to apply albeit with much smaller IGR A:B:C sizes (543:352:175 and 439:277:125 respectively) (data shown in Table 3-8). Although the datasets tested here were much larger than those used by Szafranski *et al.*, the median A:B:C ratios for *D. discoideum* and *P. falciparum* were roughly comparable. The data

for *S. cerevisiae* confirmed that, for this very small genome, the 3:2:1 A:B:C IGR spacing ratio (2:1.6:1) does not seem to apply; however, an A>B>C relationship still exists for this organism.

Table 3-8 AT-content and the IGR size and count for each A, B and C dataset for each of the 13 organisms investigated

Organism	Strain/Isolate	% AT	IGR count			Ratio of IGR count			Median size of IGR (bp)			Ratio of median size			Significant difference			Access data
			A	B	C	A	B	C	A	B	C	A	B	C	A v B	A v C	B v C	
<i>Babesia bovis</i>	Texas T2Bo	58.2	1124	1990	1032	1.1	1.9	1.0	543	352	175	3.1	2.0	1.0	Yes	Yes	Yes	GenBank 23/07/2009
<i>Cryptosporidium hominis</i>	Tu502	68.3	328	631	404	0.8	1.6	1.0	640	494	203	3.2	2.4	1.0	Yes	Yes	Yes	CryptoDB release-4.0
<i>Cryptosporidium parvum</i>	Iowa	70	994	1666	972	1.0	1.7	1.0	634	460	175	3.6	2.6	1.0	Yes	Yes	Yes	CryptoDB release-4.0
<i>Dictyostelium discoideum</i>		77.6	3312	6571	3313	1.0	2.0	1.0	825	602	241	3.4	2.5	1.0	Yes	Yes	Yes	DictyBase 26/05/2009
<i>Neospora caninum</i>	Liverpool	45.2	1694	1965	1695	1.0	1.2	1.0	3603	3899	2172	1.7	1.8	1.0	No	Yes	Yes	ToxoDB release-5.1 25/03/2009
<i>Plasmodium falciparum</i>	3D7	80.6	1405	2494	1409	1.0	1.8	1.0	1938	1385	677	2.9	2.0	1.0	Yes	Yes	Yes	PlasmoDB release-5.5
<i>Plasmodium knowlesi</i>	H strain	62.5	1320	2225	1330	1.0	1.7	1.0	2162	1592	736	2.9	2.2	1.0	Yes	Yes	Yes	PlasmoDB release-5.5
<i>Plasmodium vivax</i>	Salvador I	57.7	982	1668	944	1.0	1.8	1.0	1956	1434	643	3.0	2.2	1.0	Yes	Yes	Yes	PlasmoDB release-5.5
<i>Plasmodium yoelli</i>	17XNL	77.4	693	2679	1338	0.5	2.0	1.0	1192	578	582	2.0	1.0	1.0	Yes	Yes	No	PlasmoDB release-5.5
<i>Saccharomyces cerevisiae</i>		61.7	1424	2726	1498	1.0	1.8	1.0	485	391	238	2.0	1.6	1.0	Yes	Yes	Yes	Saccharomyces Genome DB 25/05/2009
<i>Theileria annulata</i>	Ankara clone C9	67.5	869	1856	857	1.0	2.2	1.0	439	277	125	3.5	2.2	1.0	Yes	Yes	Yes	GenBank 23/07/2009
<i>Toxoplasma gondii</i>	ME49	47.7	1134	1878	1121	1.0	1.7	1.0	2576	2437	1623	1.6	1.5	1.0	No	Yes	Yes	ToxoDB release-5.1
<i>Theileria parva</i>	Muguga	65.9	886	2052	862	1.0	2.4	1.0	376	256	154	2.4	1.7	1.0	Yes	Yes	Yes	GenBank processed 23/07/2009

This table also shows the ratio of the IGR count and the ratio of median IGR size for each A, B and C dataset for each of the 13 organisms investigated. When ratios were calculated the C value was always taken as 1.

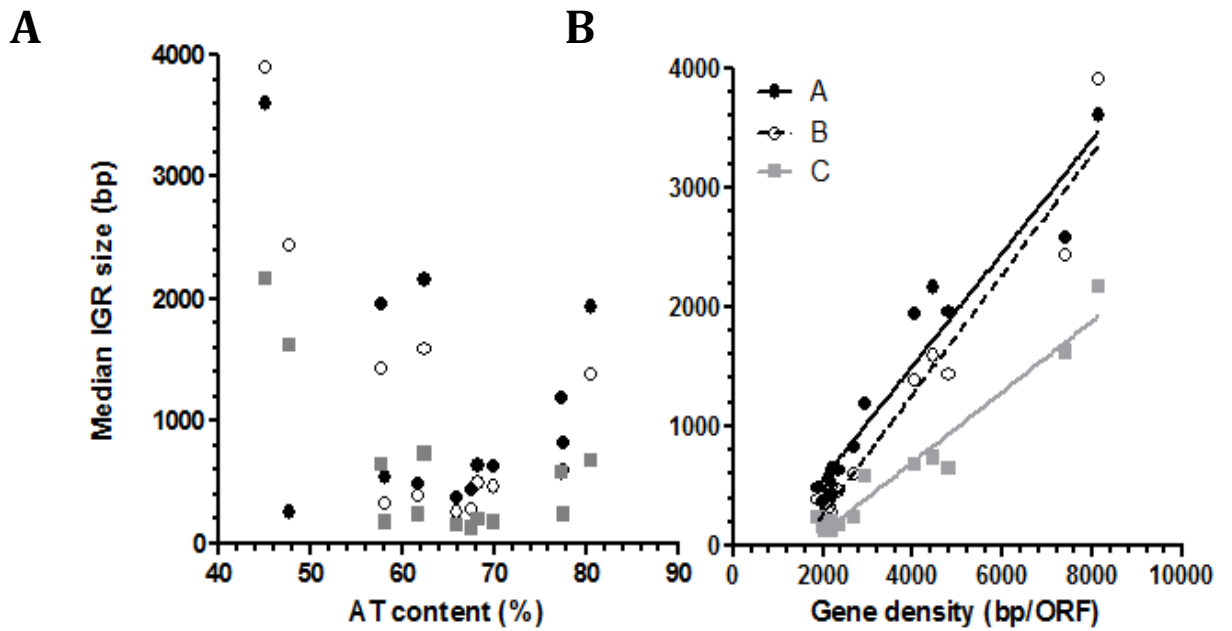


Figure 3-10 Comparison of AT content and gene density against median IGR size (bp)

(A) Plot correlating the median IGR size and AT-content for the 13 organisms investigated (AT content for each organism is shown in Table 3-8). (B) Plot correlating the median IGR size and the gene density data with linear regression lines for each of the 13 organisms investigated. Gene density was calculated by dividing the total genome size by the number of genes within each genome and these data are shown in Table 3-9. Data for Table 3-9 was taken from EuPathDB, dictyBase and the the *Saccharomyces* genome database.

Table 3-9 Comparison of the size and organisation of IGR from organisms used in this study

Organism	Gene density (kbp)
<i>Babesia bovis</i>	2.1
<i>Cryptosporidium hominis</i>	2.2
<i>Cryptosporidium parvum</i>	2.3
<i>Dictyostelium discoideum</i>	2.7
<i>Neospora caninum</i>	8.1
<i>Plasmodium falciparum</i>	4.0
<i>Plasmodium knowlesi</i>	4.4
<i>Plasmodium vivax</i>	4.8
<i>Plasmodium yoelii</i>	2.9
<i>Saccharomyces cerevisiae</i>	1.9
<i>Theileria annulata</i>	2.2
<i>Toxoplasma gondii</i>	7.4
<i>Theileria parva</i>	2.0

To explore the impact of nucleotide composition and genome density on the 3:2:1 rule scatterplots of median A, B and C IGR sizes compared with the AT content of the genome (data shown in Table 3-8) and gene density (total genome size/number of ORF) (data shown in Table 3-9) were plotted. These analyses indicate that: i) there appears to be no correlation between the sizes of different IGR and the AT content of the genome (Fig. 3-10A) but ii) linear regression analysis suggests a positive correlation between the size of IGR type and gene density (r^2 A = 0.93, B = 0.94 and C = 0.95) (Fig. 3-10B). The latter finding is perhaps not entirely unexpected – large, less dense genomes would perhaps be expected to have larger intergenic distances.

These data together suggest that, for compact genomes at least (2.0-4.8kb/ORF), where excess genomic sequence cannot afford to be ‘carried’, approximately 3 times more intergenic space appears to be required to accommodate a promoter region than a terminator region. If all intergenic sequence were functional and all intergenic sequence was mutually exclusive to its flanking gene these data could perhaps indicate an approximate 75:25% promoter:terminator region apportionment ratio.

3.7 DISCUSSION

The data presented within this chapter reveal two interesting ratios relating to the frequency and size of IGRs in the organisms investigated. In general, the frequency of A, B and C IGRs conform to a 1:2:1 ratio. Perhaps not surprising as these genomes are organized as discrete transcriptional units – but this study does represent the most complete analysis of unicellular eukaryotic organisms in general and Apicomplexans in particular to date. The second ratio, that of an approximate 3:2:1 intergenic spacing rule for types A, B and C IGRs, is perhaps the most relevant to the overall topic of this thesis. This gene spacing rule is apparent in organisms that exhibit moderately compact genomic density (2.0-4.8kb/ORF). Szafranski *et al.*, utilizing single chromosome genome sequences for *P. falciparum*, *S. cerevisiae* and *D.*

discoideum initially described this spacing rule (Szafranski *et al.*, 2005). Here, their analysis has been extended utilizing the complete genomes of these species and including a further ten organisms. These data strongly suggest that the transcriptional activity which occurs over an IGR impacts on the size of this region. That is, promoters appear to require some three times more space than terminators and this observation will be explored further in the subsequent two chapters.

The size of IGRs is, not surprisingly, directly related to genome density. As the genome size increases, the gene density decreases and intergenic distances increase. This effect is apparent even when considering the difference in total gene count across the different genus explored here. Interestingly, there is no correlation between IGR size and nucleotide content and this is perhaps best illustrated between *P. falciparum*, *P. knowlesi* and *P. vivax* which exhibit AT content of 80.6, 62.5 and 57.7% respectively yet share almost identical type A, B and C median IGR sizes. Together these data suggest that despite differences in genome size and nucleotide content, gene organization within moderately compact genomes is not random, but instead, highly organized – likely to accommodate transcriptional activity over these regions.

Comparison of intergenic space organization and size between the functionally distinct subtelomeric and chromosomal internal regions of the *P. falciparum* chromosomes demonstrates clear differences in both frequency and size. Subtelomeric regions, containing some 11.8% of all genes, demonstrate a distinct A:B:C type IGR frequency of 1:3:1 and are substantially longer with an A:B:C size ratio of 1.8:1.4:1. Most of the genes located in the subtelomeric domain are organised within families of virulence-associated genes which are under constant immune selection pressure from the human host. Antigenic diversity and regulation of clonal monallelic expression of such virulence genes plays an important role in the survival of this parasite (Scherf *et al.*, 2001; Deitsch and Hviid, 2004; Pasternak and

Dzikowski, 2009). Therefore, whilst the central parts of *P. falciparum* chromosomes are highly conserved, containing predominantly 'housekeeping genes', subtelomeric regions are highly polymorphic (Hernandez-Rivas *et al.*, 2013) - constantly increasing their repertoire of structural variants via recombination not just during sexual reproduction but also mitotically during the asexual IE cycle (Templeton, 2009; Bopp *et al.*, 2013). The terminal ends of *P. falciparum* chromosomes contain tandem GGGTT(T/C)A repeats and whilst these repeats are *Plasmodium* specific, length variance exists between species (Hernandez-Rivas *et al.*, 2013). It is thought that these telomeric regions are responsible for the observed 'tethering' of the chromosome ends into clusters at the nuclear periphery (Figueiredo *et al.*, 2002; Freitas-Junior *et al.*, 2005). Also, that this 'clustering' of telomeric ends may facilitate recombination events and, in conjunction with factors such as PfKMT1, PfSir2A, PfORC1 and PfHP1, enable telomeric heterochromatin to assemble (Freitas-Junior *et al.*, 2000; Hernandez-Rivas *et al.*, 2013). It has also been hypothesized that heterochromatin in *P. falciparum* may not be employed to silence developmentally-regulated genes, as observed in other eukaryotes, but instead, may play an important *Plasmodium*-specific role in silencing subtelomerically located virulence genes (Duffy *et al.*, 2012). Therefore, it may be reasonably considered that whilst the core genome is under selective pressure to reduce gene density, likely through reduced IGR length, perhaps driven by the demands to complete repeated rounds of replication in relatively short periods of time (e.g. during male gametocyte exflagellation), this pressure is balanced in the subtelomeric regions to ensure promotion of recombination and accommodate the repetitive sequence elements required for epigenetic regulation and peripheral nuclear positioning of chromosomal ends.

These data have demonstrated that within compact genomes, patterns of intergenic size and organisation exist. Critically, it is shown here that the median size of flanking IGR appears dependent upon the type of transcriptional activity occurring over the flanking genes - not only for *P. falciparum*, but also over a range of other Apicomplexan organisms. Thus, within a

compact genome (2.0-4.8 kb/ORF) an IGR containing two promoter regions is, on average, three times larger than an IGR containing two terminator regions. It is tempting to speculate that these data could be used to predict transcript apportionment, but this would be an oversimplification and these analyses and conclusions are based only upon median IGR values and do not take the true variance of size into account. However, it is likely that within these compact parasitic genomes, that are often under replicative pressure and where genomic 'space' is at a premium, that IGRs are kept 'concise' - only retaining information mandatory for transcriptional processes.

CHAPTER 4 A COMPARATIVE ANALYSIS OF UTR SIZE IN *P.*

FALCIPARUM

4.1 INTRODUCTION

Otto *et al.*, (2010) used RNA-Seq to improve existing gene annotation and identify splice sites within the *P. falciparum* transcriptome. Using Illumina-based massively parallel sequencing this group confirmed previous observations relating to temporally-linked mRNA accumulation during the IE cycle, although their approach appears more sensitive at detecting transcripts of lower abundance, and identified 4871 genes that were expressed during the IE cycle (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Otto *et al.*, 2010). This suggested that in excess of 80% of the genome is transcribed during this life-cycle stage to some extent. Unfortunately, the abundance of low complexity sequence regions and the high AT-content of *P. falciparum* DNA, particularly in non-coding regions, did not allow RNA-Seq reads to be allocated to regions outside of the ORF to provide the much sought after locations of transcriptional start and stop sites (Otto *et al.*, 2010).

Ideally, any approach directed towards an understanding of the molecular mechanisms that govern the control of gene expression would be best supported by clear delineation of critical transcriptional landmarks – particularly transcriptional start and stop sites. Without these landmarks, putative *cis*-acting regulatory sequences that initiate and control transcription are notoriously difficult to predict from the extremely AT-rich flanking sequences of *P. falciparum*. Whilst numerous putative regulatory motifs have been predicted using various bioinformatic approaches (Elemento *et al.*, 2007; Young *et al.*, 2008; Wu *et al.*, 2008), relatively few have been experimentally validated and even fewer assessed for their potential roles in the control of gene expression. Several approaches have been used to identify transcriptional landmarks. These include: Expressed sequence tags (EST), bioinformatic

prediction of core promoter regions, physical mapping approaches such as RNase protection and functional studies utilizing promoter/terminator deletions in reporter gene assays (reviewed in Horrocks *et al.*, 2009).

Expressed sequence tag (EST) data are available through various sources such as dBEST (<http://www.ncbi.nlm.nih.gov/dbEST>), Full-Malaria (<http://fullmal.hgc.jp>) and PlasmoDB (<http://plasmodb.org/plasmo>). Conventionally cDNAs were synthesised using a dT primer that hybridised to the mRNA PolyA tail. Unfortunately, this heavily selected for the 3' end of the transcript and the use of reverse transcriptase with low processivity over AT rich sequences, particularly homopolymeric poly dA:dT tracts, resulted in a series of truncated products. This issue was addressed by Watanabe *et al.*, (2001, 2004) who utilised RNA Ligation mediated rapid amplification of cDNA ends (RLM-RACE) methodology to construct a full-length cDNA library. By the introduction of a sequence tag, in the form of an oligo cap at the 5' end of the transcript (replacing the 5' 5mG transcript cap), this allowed for both 3' and 5' selection of transcripts (Watanabe *et al.*, 2001; Watanabe *et al.*, 2004). These data are now available in the Full-Malaria database, unfortunately, coverage is not complete and data is only available for ~27% of *P. falciparum* genes (Watanabe *et al.*, 2007). These data suggest that multiple transcription start sites exist and that they are typically located within the 150-450bps region upstream of the ORF (Watanabe *et al.*, 2002).

In 2008 Brick *et al.*, utilized a bioinformatics algorithm to search for core promoters in the 5' flanking sequences of *P. falciparum* ORF. The Malaria Promoter Predictor (MAPP) was based on 33 biophysical properties of DNA associated with the core promoter structure and trained in part on the Watanabe *et al.* cDNA dataset. These data generally supported the predicted transcription start site data available from RLM-RACE, although core promoter distribution did appear to extend slightly further upstream of the ORF (in the region of 400-900bp) (Brick *et al.*, 2008).

Very few transcription start site (TSS) mapping studies or proper functional assays have been carried out in *P. falciparum* (summary Table 1-1) (Horrocks *et al.*, 2009; Wong *et al.*, 2011), but these tend to suggest that TSS are somewhere in the region of 400-1900bp upstream of the ORF and that there are generally only one or two primary TSS – somewhat at odds with the RLM-RACE data. The predicted presence of multiple transcription start sites is also at odds with what is typically observed from northern blot data, which tends to show one or two principle transcripts, and that these transcripts are much larger than the ORF they transcribe. Despite this disparity between EST and northern blot data regarding the size of a transcriptional unit, no systematic comparison of these datasets exists.

The data presented within this chapter attempts a systematic comparison of northern blot and EST data for a cohort of 105 genes for which northern blot data was either secured from the literature or generated *de novo* for this study. Utilizing the relative strengths of each method an attempt is made to describe the size and apportionment of the untranslated region (UTR) of the transcript to better understand where the key transcriptional start and terminator landmarks lie.

4.2 EXPLORING THE RELATIONSHIP BETWEEN ORF AND TRANSCRIPT SIZE

4.2.1 COLLECTION OF NORTHERN BLOT DATA

A cohort of Northern blot data was collected from 105 ORF. Of these, 62 were gathered during a review of the published literature with the remaining 43 Northern blots carried out during this and other studies in our laboratory (Table 4-1).

A review of the available literature was carried out By searching PubMed using terms such as; ‘falciparum’, ‘malaria’, ‘transcript’, ‘northern blot’, ‘transcription’ separately, or in different combinations. Identified manuscripts were sourced and the following criteria applied to the

use of northern blot data; (i) the transcript size needed to be specifically stated in the text OR an image of a northern blot shown with sufficient marker information to estimate the size of the transcript, (ii) the gene of interest needed to be specifically stated and (iii) transcript data from large multigene families (*e.g. var*) were excluded owing to ambiguous annotation issues. This search provided data on transcript size for 62 genes.

Forty-three northern blots were undertaken *de novo* in the laboratory using the protocol described in the materials and methods (Kyes *et al.*, 2000) - thirty-seven by myself; the remainder being unpublished data from the laboratory. To determine the size of the transcript mixed IE stage or trophozoite stage RNA was isolated, size fractionated, blotted and probed using an $\alpha^{32}\text{P}$ -dADP random-primer labelled PCR fragment from the gene of interest. Oligonucleotides used to generate the products are shown in Table 1-1 (Materials and Methods). Products ranged from 412-818bp in size and were designed to a single exon (example northern blots are shown in Fig. 4-1). Combining the two sets of data provided a cohort of 105 genes (Table 4-1).

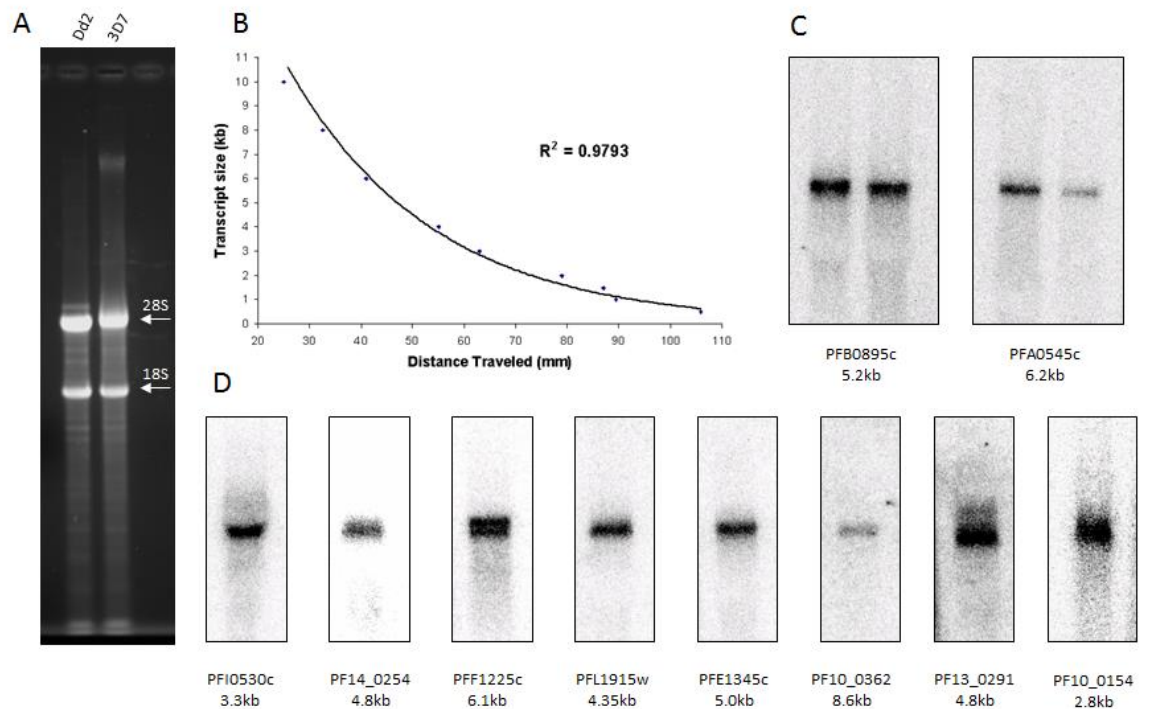


Figure 4-1 Example northern blot data generated *de novo* for this study

(A) Size-fractionated total RNA (5µg per lane) from Dd2 and 3D7 clones, the position of the two largest ribosomal bands is indicated. (B) An example calibration graph used to determine transcript size based on the migration of RNA markers (Gibco-BRL). Note the length of the poly A tail was not taken into account when sizing transcripts (C) Example northern blot data from blot of gel shown in (A); the gene investigated and transcript size estimated are shown below. (D) Example northern blot data from mixed stage RNA from 3D7; the gene investigated and transcript size estimated are shown below.

Table 4-1 Cohort of 105 ORF from *P. falciparum* for which northern blot data was collated

(overpage) The PlasmoDB annotation and unique identifier (note historical identifier used as relates to published data prior to adoption of new naming convention), ORF size and reported size of transcript from Northern blot and the number of exons in the ORF are indicated. ¹ IDC profile - Peak temporal window of transcription was defined from the malaria IDC strain comparison database (<http://malaria.ucsf.edu/comparison/index.php>) where R, Ring was 1-16 hour post-infection (hpi), T, Trophozoite 17-32hpi and S, Schizont 32-48hpi. Note, several ORF are either not transcribed during the IDC or unambiguous microarray data was not available. ²ORF orientation is defined into three subgroups (A,B,C) that relate to the orientation of flanking ORF: type A, flanking ORF are both transcribed away from ORF in question, type B, one flanking ORF is transcribed towards the ORF in question and one away and, type C, both flanking ORF are transcribed away. The final column provides a reference for the source of data or indicates it is our own unpublished data (this study).

PlasmoDB Annotation	Gene ID	ORF kb	Transcript kb	No. exons	T'scription ¹	Orientation ²	Reference
Chloroquine resistance transporter	MAL7P1.27	1.275	4.20	13	R	A	Waller <i>et al.</i> , 2003
MSP7-like protein	PF13_0193	0.897	1.70	1	R	B	Mello <i>et al.</i> , 2002
Actin II	PF14_0124	1.131	1.90	2	R	B	Wesseling <i>et al.</i> , 1989
Conserved protein, unknown function	PFB0115w	3.579	5.00	1	R	B	Lanzer <i>et al.</i> , 1994
Adenylosuccinate lyase, putative	PFB0295w	1.416	3.00	1	R	B	Kyes <i>et al.</i> , 2002
DNA-directed RNA polymerase II, putative	PFC0805w	7.374	9.00	1	R	C	Li <i>et al.</i> , 1989
Multidrug resistance protein(PfMDR)	PFE1150w	4.26	7.50	1	R	C	Myrick <i>et al.</i> , 2003, Foote <i>et al.</i> , 1989, Wilson <i>et al.</i> , 1989, Volkman <i>et al.</i> , 1993
Early transcribed membrane protein 5	PFE1590w	0.546	2.70	1	R	C	Spielmann and Beck, 2000
Hexokinase	PFF1155w	1.482	3.00	1	R	B	Olafsson <i>et al.</i> , 1992
Histone deacetylase	PFI1260c	1.35	2.80	1	R	B	Joshi <i>et al.</i> , 1999
MSP7-like protein	MAL13P1.174	0.846	1.50	1	S	B	Mello <i>et al.</i> , 2002
Drug/metabolite exporter, drug/metabolite transporter	PF07_0064	1.305	2.90	1	S	B	This study
Zinc transporter, putative	PF07_0065	1.671	5.10	1	S	A	This study
Transcriptional activator ADA2, putative	PF10_0143	7.737	8.50	1	S	B	Fan <i>et al.</i> , 2004a
Ribonucleotide reductase small subunit, putative	PF10_0154	1.008	2.80	5	S	A	This study
Microtubule-associated protein 1 light chain 3, putative	PF10_0193	0.375	1.80	1	S	B	This study
DNA polymerase zeta catalytic subunit, putative	PF10_0362	6.723	8.60	3	S	B	This study
Early transcribed membrane protein 11.1	PF11_0039	0.276	1.30	1	S	B	Spielmann and Beck, 2000
Deoxyuridine 5'-triphosphate nucleotidohydrolase, putative	PF11_0282	0.522	1.40	1	S	A	This study
Aquaglyceroporin PfAQP	PF11_0338	0.777	1.80	1	S	B	Hansen <i>et al.</i> , 2002
Casein kinase 1 PfCK1	PF11_0377	0.972	2.40	8	S	A	Barick <i>et al.</i> , 1997
Conserved Plasmodium protein, unknown function	PF11_0425	1.062	3.40	5	S	A	This study
Hypothetical protein, Aos1	PF11_0457	1.017	2.60	1	S	B	This study
Dynammin-like protein	PF11_0465	2.514	3.00	4	S	A	Li <i>et al.</i> , 2004
Merozoite adhesive erythrocyte binding protein maeb1	PF11_0486	6.168	8.00	5	S	C	Balu <i>et al.</i> , 2008
Pf gamete antigen Pfg 27/25	PF13_0011	0.654	2.50	1	S	B	Alano <i>et al.</i> , 1996
Conserved Plasmodium protein, unknown function	PF13_0199	4.83	8.00	3	S	B	This study
Conserved Plasmodium protein, unknown function	PF13_0200	1.812	3.40	1	S	B	This study

Proliferating cell nuclear antigen	PF13_0328	0.825	1.95	1	S	B	This study
Uracil-DNA glycosylase, putative	PF14_0148	0.969	2.80	1	S	A	This study
Conserved Plasmodium protein, unknown function	PF14_0149	1.773	3.30	3	S	A	This study
Minichromosome maintenance complex subunit	PF14_0177	2.916	4.80	2	S	C	This study
Serine/threonine protein phosphatase	PF14_0224	2.88	3.60	16	S	B	Dobson <i>et al.</i> , 2001
ATP-specific succinyl-CoA synthetase beta subunit, putative	PF14_0295	1.389	2.25	1	S	A	This study
DNA topoisomerase II, putative	PF14_0316	4.419	6.00	4	S	A	Cheesman <i>et al.</i> , 1998
Calmodulin	PF14_0323	0.45	1.50	2	S	B	Kyes <i>et al.</i> , 2000
Fructose-bisphosphate aldolase	PF14_0425	1.11	2.40	2	S	B	Knapp <i>et al.</i> , 1990
DNA polymerase alpha subunit, putative	PF14_0602	1.62	3.20	1	S	B	This study
Conserved Plasmodium protein, unknown function	PFA0285c	2.499	4.95	2	S	B	This study
DNA binding protein, putative	PFA0290w	1.347	2.80	1	S	B	This study
Replication factor C protein	PFA0545c	3.504	6.20	1	S	A	This study
Merozoite surface protein 2 precursor	PFB0300c	0.819	2.00	1	S	C	Kyes <i>et al.</i> , 2002
PfPCK calcium dependent protein kinase 1	PFB0815w	1.575	3.20	5	S	C	Zhao <i>et al.</i> , 1993
Replication factor C, subunit 2	PFB0840w	0.993	2.40	1	S	A	This study
Hypothetical protein, conserved	PFC0090w	0.774	2.30	2	S	B	Spielmann and Beck, 2000
DNA polymerase alpha	PFD0590c	5.739	7.00	5	S	A	White <i>et al.</i> , 1993
Conserved Apicomplexan protein, unknown function	PFD0595w	2.337	4.45	1	S	B	This study
Phosphoglycerate mutase, putative	PFD0660w	0.888	2.40	1	S	C	This study
Cdc2-related protein kinase 1	PFD0865c	2.1	3.50	1	S	B	Doerig <i>et al.</i> , 1995
Alpha tubulin II	PFD1050w	1.353	2.80	3	S	A	Delves <i>et al.</i> , 1990
Asparagine rich protein PfAARP	PFD1105w	0.654	1.80	1	S	A	Wickramarachchi <i>et al.</i> , 2008
Topoisomerase I	PFE0520c	2.52	3.80	1	S	B	Tosh and Kilbey, 1995, Tosh <i>et al.</i> , 1999
Minichromosome maintenance complex subunit, putative	PFE1345c	2.889	5.00	2	S	C	This study
Conserved Plasmodium protein, unknown function	PFI0540w	3.498	5.00	2	S	B	This study
Merozoite surface protein 1 precursor MSP-1	PFI1475w	5.163	8.00	1	S	A	Kyes <i>et al.</i> , 2000
Transcription factor with AP2 domains, putative	PFI1665w	0.603	1.50	1	S	C	Spielmann and Beck, 2000
Ubiquitin activating enzyme, putative	PFL1790w	2.061	3.90	2	S	B	This study
Actin I	PFL2215w	1.131	2.50	1	S	B	Horrocks <i>et al.</i> , 1996, Cheesman <i>et al.</i> , 1998, Horrocks <i>et al.</i> , 2002, Wesseling <i>et al.</i> , 1989
Pf protein kinase 6	MAL13P1.185	0.918	2.00	8	T	A	Bracchi-Ricard <i>et al.</i> , 2000
Histone acetyltransferase GCNS, putative	PF08_0034	4.398	5.40	4	T	A	Fan <i>et al.</i> , 2004b
1-cys peroxidoxin	PF08_0131	0.663	2.20	1	T	B	Spielmann and Beck, 2000

Acyl CoA binding protein, isoform 2 ACBP2	PF10_0016	0.273	1.70	1	T	B	Spielmann and Beck, 2000
Tubulin beta chain, putative	PF10_0084	1.338	2.70	3	T	C	Delves <i>et al.</i> , 1990
Hsp60	PF10_0153	1.743	2.60	2	T	B	Syin and Goldman, 1996
Serine/threonine protein kinase FIKK family	PF10_0160	1.83	3.80	3	T	B	This study
DNA polymerase delta catalytic subunit	PF10_0165	3.285	5.70	1	T	B	Lanzer <i>et al.</i> , 1992a, Horrocks <i>et al.</i> , 1996, Cheesman <i>et al.</i> , 1998, Horrocks <i>et al.</i> , 2002
Histone H4, putative	PF11_0061	0.312	2.00	1	T	B	Przyborski <i>et al.</i> , 2003
Replication factor C subunit 5, putative	PF11_0117	1.05	3.00	1	T	C	This study
ThiF family protein, putative	PF11_0271	3.951	4.70	2	T	B	This study
DNA-directed RNA polymerase III largest subunit	PF13_0150	7.071	8.50	5	T	B	Li <i>et al.</i> , 1991
Minichromosome maintenance complex subunit	PF13_0291	2.79	4.80	1	T	C	This study
Ribonucleotide reductase small subunit	PF14_0053	1.05	2.50	1	T	C	This study
Plasmepsin III HAP protein	PF14_0078	1.356	2.80	1	T	C	Berry <i>et al.</i> , 1999
Cytidine diphosphate-diacylglycerol synthase	PF14_0097	2.004	3.50	1	T	C	Martin <i>et al.</i> , 2000, Osta <i>et al.</i> , 2002
DNA mismatch repair protein Msh2p, putative	PF14_0254	2.436	4.80	1	T	A	This study
Replication factor C subunit 1, putative	PFB0895c	2.715	5.20	1	T	B	This study
DNA polymerase epsilon subunit b, putative	PFC0340w	1.497	3.20	1	T	A	This study
Ubiquitin-like protein SUMO	PFE0285C	0.303	0.80	2	T	B	This study
DNA polymerase 1	PF11225c	4.335	6.10	1	T	C	This study
PfRab7, GTPase	PFI0155c	0.621	2.60	7	T	A	Spielmann and Beck, 2000
Alpha tubulin	PFI0180w	1.362	2.90	3	T	B	Delves <i>et al.</i> , 1990
Replication factor A-related protein, putative	PFI0235w	1.455	2.90	2	T	B	This study
DNA primase, large subunit, putative	PFI0530c	1.62	3.30	2	T	B	This study
GINS complex subunit Psf3, putative	PFI0725c	0.627	3.60	1	T	B	This study
Ubiquitin conjugating enzyme, putative	PFI0740c	0.48	2.20	4	T	A	This study
Phosphoglycerate kinase	PFI1105w	1.251	2.10	1	T	B	Hicks <i>et al.</i> , 1991
Thioredoxin reductase	PFI1170c	1.854	3.20	1	T	B	Krnajski <i>et al.</i> , 2002
Proliferating cell nuclear antigen 2, putative	PFL1285c	0.795	2.35	2	T	B	This study
Chaperonin, cpn60	PFL1545c	2.157	4.00	2	T	A	Holloway <i>et al.</i> , 1994
DNA polymerase epsilon subunit b, putative	PFL1655c	1.875	3.50	2	T	C	This study
DNA gyrase subunit b, putative	PFL1915w	3.021	4.35	1	T	B	This study
Replication factor C, subunit 4	PFL2005w	1.011	2.60	2	T	B	This study
Plasmodium exported protein, unknown function	MAL7P1.170	0.882	2.40	2		A	Spielmann and Beck, 2000
Heat shock 70kDa protein, putative hsp70	MAL7P1.228	1.986	4.00	2		C	Przyborski <i>et al.</i> , 2003
Histidine rich protein II	MAL7P1.231	0.918	2.10	2		A	Horrocks <i>et al.</i> , 2002, Wellems and Howard, 1986

Glycophorin-binding protein 130 precursor	PF10_0159	2.475	6.60	2	C	Horrocks <i>et al.</i> , 1996, Cheesman <i>et al.</i> , 1998, Lanzer <i>et al.</i> , 1992a
MSP7-like protein	PF13_0196	1.143	1.90	1	B	Mello <i>et al.</i> , 2002
Membrane-associated histidine rich protein 2 MARHP2	PF13_0276	0.414	2.60	2	B	Spielmann and Beck, 2000
Elongation factor 1 alpha	PF13_0304	1.332	2.40	1	A	Waller <i>et al.</i> , 2003
Mitogen-activated protein kinase 1, PfMAP1	PF14_0294	2.745	3.70	1	A	Doerig <i>et al.</i> , 1996
Small subunit DNA primase	PF14_0366	1.359	2.10	16	B	Prasartkaew <i>et al.</i> , 1996
Knob associated histidine-rich protein kahrp	PFB0100c	1.965	4.20	2	B	Lanzer <i>et al.</i> , 1992b, Lanzer <i>et al.</i> , 1994
Hsp40	PFD0462w	2.019	3.10	1	B	Watanabe <i>et al.</i> , 1997
Nucleosome assembly protein	PFI0930c	0.81	1.40	3	B	Dobson <i>et al.</i> , 2003
Pf polyubiquitin PfpUB	PFL0585w	1.146	2.30	2	C	Horrocks and Newbold, 2000

A full reference list for all northern blot references secured from the literature is shown in Appendix D-1

4.2.2 ANALYSIS OF UTR SIZES DERIVED FROM NORTHERN BLOT DATA

Using the predicted sizes of the ORF (with intronic sequence removed) and the transcript size from the northern blots, the length of the UTR within each transcript was estimated. The size of the predicted UTR from these 105 transcripts revealed a distribution between 486 and 4125 bases (Fig. 4-2A, median 1518, interquartile range 1150-1844 bases). There was insufficient data to demonstrate a normal distribution, although there is clearly an evolving pattern of mono-modal distribution with 72% of all UTR sizes falling between 800-1800 bases. Comparing the UTR size against the size of their respective ORF reveals no significant correlation (Figure 2B, $R^2=0.04$), i.e. the size of the ORF does not appear to affect the size of the UTR arranged around it in the transcript. Given the apparent restricted distribution of the majority of UTR sizes, it was not surprising to find a good correlation between the sizes of the ORF and the whole transcript (Figure 2C, $R^2=0.88$), with a slope close to one (1.07 ± 0.04) and a y-intercept of 1444 ± 99 bases (close to the median distribution of 1518 bases).

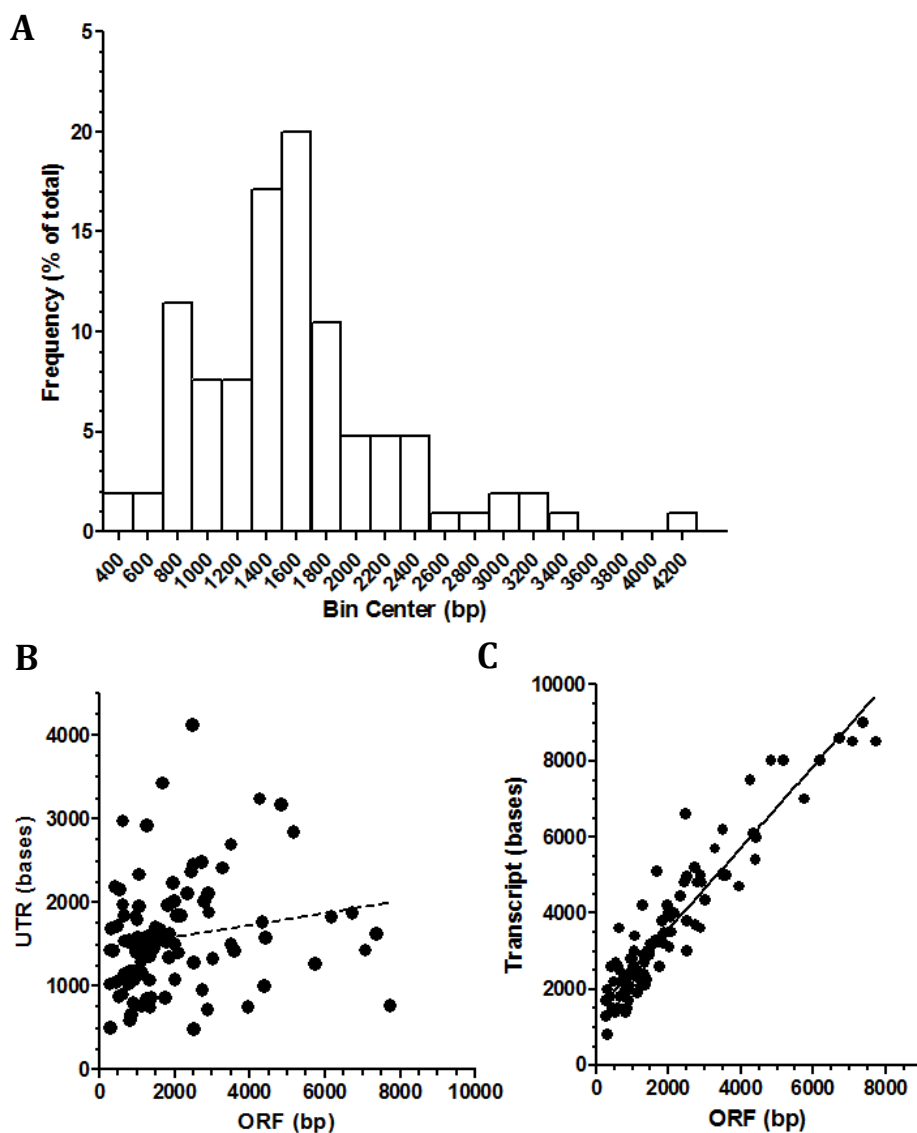


Figure 4-2 UTR lengths from northern blot data

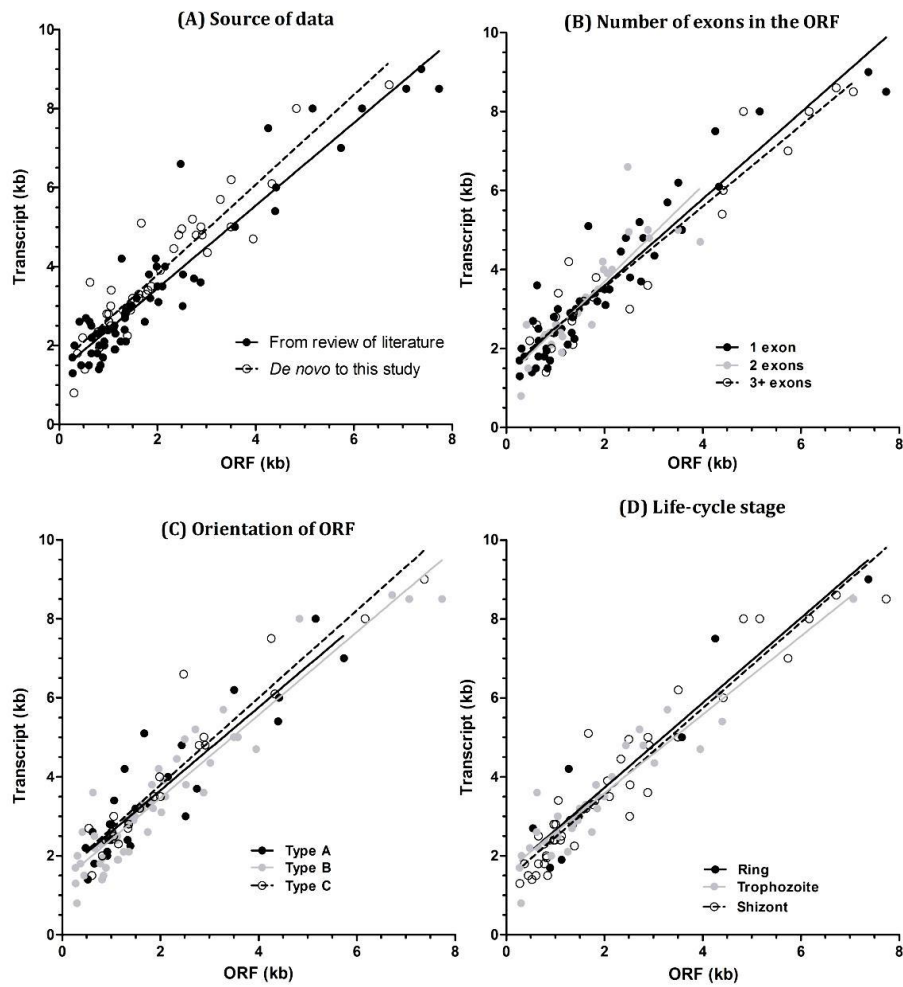
(A) Frequency distribution of the UTR (200 base bins) (B) Scattergram and correlation between the predicted size of the UTR (bases) and the ORF (bp) ($R^2 = 0.04$) and (C) Scattergram and correlation between the size of the ORF (bases) and transcript (bp) ($R^2 = 0.88$).

Next, the relationship between transcript and ORF sizes was further explored to understand whether this relationship remained when the data was granulated using a series of different criteria. These criteria were; (i) the source of the transcript data (published data *vs. de novo* data), (ii) the organisation of the ORF in terms of the number of exons, (iii) the organisation

of the ORF in terms of its orientation with respect to flanking genes and (iv) the principle morphological stage during IE development at which the peak of steady-state transcription occurs (ring vs. trophozoite vs. schizont - see Appendix D2). Plotting the transcript vs. the ORF size using these criteria revealed no significant differences between the correlation, slope and y-intercept of the regression analyses, irrespective of the criteria applied (Fig. 4-3). Thus, the source of transcript data, the organisational features of the genes explored and the stage of peak mRNA accumulation during the IE cycle do not appear to contribute to the size of the UTR explored here.

Figure 4-3 Linear regression analysis of the transcript and ORF size

(overpage) Data has been categorised according to; (A) source, (B) gene structure based on exon number, (C) orientation of the ORF in question with respect to the two flanking ORF and (D) the major morphological stage of IDC where peak temporal transcript steady state levels are determined (see Appendix D-2 for IDC expression data). A key located on each graph indicates the variables used to divide the cohort of data. The table reports the results of all linear regressions as number of ORF in subset (n=). Analysis of covariance in each case revealed no significant difference between slopes and y-intercept.



Source	n=	Slope	Y-intercept (bp)	R ²
All data	105	1.07 ± 0.04	1444 ± 99	0.88
Graph A				
Published data	62	1.05 ± 0.05	1351 ± 123	0.90
This study	43	1.14 ± 0.07	1518 ± 158	0.88
Graph B				
1 exon	56	1.10 ± 0.05	1403 ± 127	0.89
2 exons	25	1.22 ± 0.15	1263 ± 286	0.75
3 or more exons	24	1.02 ± 0.07	1496 ± 218	0.92
Graph C				
Type A	28	1.05 ± 0.09	1557 ± 225	0.83
Type B	55	1.05 ± 0.05	1362 ± 119	0.90
Type C	22	1.11 ± 0.08	1579 ± 240	0.90
Graph D				
Ring	10	1.07 ± 0.13	1595 ± 407	0.89
Trophozoite	34	0.99 ± 0.06	1607 ± 147	0.89
Schizont	78	1.09 ± 0.05	1383 ± 136	0.91

4.3 ANALYSIS OF UTR SIZES DERIVED FROM EST DATA

Available 5' and 3' EST data was secured from the Full-Malaria (Watanabe *et al.*, 2004) and dbEST databases for the cohort of 105 genes for which northern blot data were available (Table 4-2). For 44 of the 105 genes in this cohort, both 5' and 3' EST data was available. Summing these two values gave a predicted UTR length from EST data which we used to (i) compare to the northern blot UTR data and (ii) explore the apportionment of UTR within the transcript. The size of these EST UTR sizes ranged between 80-952bases, with a median value of 512 bases and an interquartile range of 351 to 630 bases (Fig. 4-4A). A mono-modal distribution was once again apparent, although these datasets were, once again, too small to demonstrate a normal distribution. The UTR sizes from EST were considerably smaller than those established from the northern blot cohort. A scatterplot analysis clearly indicates no correlation between the UTR sizes as determined from EST and northern blot datasets for these 44 genes (Fig. 4-4B, $R^2 = 0.02$) (Fig. 4-4B).

Table 4-2 Comparative data between UTR sizes determined from northern blot and EST sources

(overpage) The table shows the; PlasmoDB identifier, ORF size (bp), transcript size from northern blot (bases). ¹putative UTR from northern which was established by subtracting the size of the ORF from the transcript, the most distal 5' EST value, the most distal 3' EST value and ²the total estimated UTR size from the EST data by summing the 5' and 3' EST values, the 5' and 3' EST data values expressed as a % of the total EST UTR, the 19 ORF for which a consensus canonical polyadeylation site motif could be identified (yes/no) and ³the putative 3' UTR apportionment (%) as a fraction of the UTR estimated from the northern blot.

PlasmoDB Annotation	Gene ID	Open Reading Frame (bp)	Transcript from Northern blot (bp)	UTR ¹ from Northern blot (bp)	5' EST (bp)	3' EST (bp)	UTR ² from EST (bp)	%5' UTR	%3' UTR	PolyA consensus signal	3' UTR ³ as % of Northern blot UTR
MSP7-like protein	MAL13P1.174	846	1500	654							
Pf protein kinase 6	MAL13P1.185	918	2000	1082	108	5	113	95.6	4.4	no	0.5
Plasmodium exported protein, unknown function	MAL7P1.170	882	2400	1518	408						
Heat shock 70kDa protein, putative (hsp70)	MAL7P1.228	1986	4000	2014							
Histidine rich protein II	MAL7P1.231	918	2100	1182	367	187	554	66.2	33.8	yes	15.8
Chloroquine resistance transporter	MAL7P1.27	1275	4200	2925							
Drug/metabolite exporter, drug/metabolite transporter	PF07_0064	1305	2900	1595		26					
Zinc transporter, putative	PF07_0065	1671	5100	3429	323	33	356	90.7	9.3	no	1.0
Hisone acetyltransferase (GCNS), putative	PF08_0034	4398	5400	1002	349	41	390	89.5	10.5	no	4.1
1-cys peroxidoxin	PF08_0131	663	2200	1537	54	394	448	12.1	87.9	yes	25.6
Acyl CoA binding protein, isoform 2 ACBP2	PF10_0016	273	1700	1427	410	359	769	53.3	46.7	yes	25.2
Tubulin beta chain, putative	PF10_0084	1338	2700	1362	451	106	557	81.0	19.0	no	7.8
Transcriptional activator ADA2, putative	PF10_0143	7737	8500	763							
Hsp60	PF10_0153	1743	2600	857	270	239	509	53.0	47.0	yes	27.9
Ribonucleotide reductase small subunit, putative	PF10_0154	1008	2800	1792	199						
Glycophorin-binding protein 130 precursor	PF10_0159	2475	6600	4125	693	90	783	88.5	11.5	no	2.2
Serine/threonine protein kinase FIKK family	PF10_0160	1830	3800	1970	198	154	352	56.3	43.8	no	7.8
DNA polymerase delta catalytic subunit	PF10_0165	3285	5700	2415	369						
Microtubule-associated protein 1 light chain 3, putative	PF10_0193	375	1800	1425	19						
DNA polymerase zeta catalytic subunit, putative	PF10_0362	6723	8600	1877		63					
Early transcribed membrane protein 11.1, etramp 11.1	PF11_0039	276	1300	1024							
Histone H4, putative	PF11_0061	312	2000	1688	328	185	513	63.9	36.1	no	11.0
Replication factor C subunit 5, putative	PF11_0117	1050	3000	1950	184						
ThiF family protein, putative	PF11_0271	3951	4700	749							
Deoxyuridine 5'-triphosphate nucleotidohydrolase, putative	PF11_0282	522	1400	878	466	68	534	87.3	12.7	no	7.7
Aquaglyceroporin (PfAQP)	PF11_0338	777	1800	1023	385	236	621	62.0	38.0	yes	23.1
Casein kinase 1 (PfCK1)	PF11_0377	972	2400	1428	589	310	899	65.5	34.5	yes	21.7
Conserved Plasmodium protein, unknown function	PF11_0425	1062	3400	2338	181						

Hypothetical protein, (Aos1)	PF11_0457	1017	2600	1583								
Dynamain-like protein	PF11_0465	2514	3000	486		44						
Merozoite adhesive erythrocyte binding protein (maebl)	PF11_0486	6168	8000	1832		37						
Pf gamete antigen 27/25 (Pfg 27/25)	PF13_0011	654	2500	1846	429	395	824	52.1	47.9	yes	21.4	
DNA-directed RNA polymerase III largest subunit	PF13_0150	7071	8500	1429								
MSP7-like protein	PF13_0193	897	1700	803		467						
MSP7-like protein	PF13_0196	1143	1900	757								
Conserved Plasmodium protein, unknown function	PF13_0199	4830	8000	3170	393							
Conserved Plasmodium protein, unknown function	PF13_0200	1812	3400	1588								
Membrane-associated histidine rich protein 2 (MARHP2)	PF13_0276	414	2600	2186								
Minichromosome maintenance (MCM) complex subunit	PF13_0291	2790	4800	2010	462							
Elongation factor 1 alpha	PF13_0304	1332	2400	1068	527	103	630	83.7	16.3	no	9.6	
Proliferating cell nuclear antigen	PF13_0328	825	1950	1125	429	241	670	64.0	36.0	yes	21.4	
Ribonucleotide reductase small subunit	PF14_0053	1050	2500	1450	10	217	227	4.4	95.6	yes	15.0	
Plasmepsin III HAP protein	PF14_0078	1356	2800	1444	530	72	602	88.0	12.0	no	5.0	
Cytidine diphosphate-diacylglycerol synthase	PF14_0097	2004	3500	1496		26						
Actin II	PF14_0124	1131	1900	769	503	177	680	74.0	26.0	yes	23.0	
Uracil-DNA glycosylase, putative	PF14_0148	969	2800	1831		47						
Conserved Plasmodium protein, unknown function	PF14_0149	1773	3300	1527		96						
Minichromosome maintenance (MCM) complex subunit	PF14_0177	2916	4800	1884	291	84	375	77.6	22.4	no	4.5	
Serine/threonine protein phosphatase	PF14_0224	2880	3600	720		134						
DNA mismatch repair protein Msh2p, putative	PF14_0254	2436	4800	2364		334						
Mitogen-activated protein kinase 1, PfMAP1	PF14_0294	2745	3700	955	29							
ATP-specific succinyl-CoA synthetase beta subunit, putative	PF14_0295	1389	2250	861		96						
DNA topoisomerase II, putative	PF14_0316	4419	6000	1581								
Calmodulin	PF14_0323	450	1500	1050	651	288	939	69.3	30.7	yes	27.4	
Small subunit DNA primase	PF14_0366	1359	2100	741	152	199	351	43.3	56.7	yes	26.9	
Fructose-bisphosphate aldolase	PF14_0425	1110	2400	1290	409	132	541	75.6	24.4	no	10.2	
DNA polymerase alpha subunit, putative	PF14_0602	1620	3200	1580								
Conserved Plasmodium protein, unknown function	PFA0285c	2499	4950	2451								
DNA binding protein, putative	PFA0290w	1347	2800	1453	14							

Replication factor C protein	PFA0545c	3504	6200	2696	115							
Knob associated histidine-rich protein (kahrp)	PFB0100c	1965	4200	2235	605	347	952	63.6	36.4	yes	15.5	
Conserved protein, unknown function	PFB0115w	3579	5000	1421	53							
Adenylosuccinate lyase, putative	PFB0295w	1416	3000	1584	204							
Merozoite surface protein 2 precursor (MSP-2)	PFB0300c	819	2000	1181	383	354	737	52.0	48.0	yes	30.0	
PfPCK calcium dependent protein kinase 1	PFB0815w	1575	3200	1625	36	226	262	13.7	86.3	yes	13.9	
Replication factor C, subunit 2	PFB0840w	993	2400	1407	297	24	321	92.5	7.5	no	1.7	
Replication factor C subunit 1, putative	PFB0895c	2715	5200	2485								
Hypothetical protein, conserved	PFC0090w	774	2300	1526	628							
DNA polymerase epsilon subunit b, putative	PFC0340w	1497	3200	1703	171							
DNA-directed RNA polymerase II, putative	PFC0805w	7374	9000	1626		39						
Hsp40	PFD0462w	2019	3100	1081	437							
DNA polymerase alpha	PFD0590c	5739	7000	1261								
Conserved Apicomplexan protein, unknown function	PFD0595w	2337	4450	2113								
Phosphoglycerate mutase, putative	PFD0660w	888	2400	1512	227	276	503	45.1	54.9	no	18.3	
Cdc2-related protein kinase 1	PFD0865c	2100	3500	1400		108						
Alpha tubulin II	PFD1050w	1353	2800	1447	394							
Asparagine rich protein (PfAARP)	PFD1105w	654	1800	1146	76	233	309	24.6	75.4	no	20.3	
Ubiquitin-like protein (Sumo)	PFE0285C	303	800	497		131						
Topoisomerase I	PFE0520c	2520	3800	1280	451							
Multidrug resistance protein	PFE1150w	4260	7500	3240	41							
Minichromosome maintenance (MCM) complex subunit, putative	PFE1345c	2889	5000	2111	454	105	559	81.2	18.8	no	5.0	
Early transcribed membrane protein 5, etramp 5	PFE1590w	546	2700	2154	463	423	886	52.3	47.7	yes	19.6	
Hexokinase	PFF1155w	1482	3000	1518		8						
DNA polymerase 1	PFF1225c	4335	6100	1765	96	170	266	36.1	63.9	no	9.6	
PfRab7, GTPase	PFI0155c	621	2600	1979	200	283	483	41.4	58.6	yes	14.3	
Alpha tubulin	PFI0180w	1362	2900	1538	394							
Replication factor A-related protein, putative	PFI0235w	1455	2900	1445	201	311	512	39.3	60.7	yes	21.5	
DNA primase, large subunit, putative	PFI0530c	1620	3300	1680	239	112	351	68.1	31.9	no	6.7	
Conserved Plasmodium protein, unknown function	PFI0540w	3498	5000	1502								
GINS complex subunit Psf3, putative	PFI0725c	627	3600	2973								

Ubiquitine conjugating enzyme, putative	PFI0740c	480	2200	1720	519	24	543	95.6	4.4	no	1.4
Nucleosome assembly protein	PFI0930c	810	1400	590		163					
Phosphoglycerate kinase	PFI1105w	1251	2100	849	248	207	455	54.5	45.5	yes	24.4
Thioredoxin reductase	PFI1170c	1854	3200	1346	41	39	80	51.3	48.8	no	2.9
Histone deacetylase	PFI1260c	1350	2800	1450							
Merozoite surface protein 1 precursor (MSP-1)	PFI1475w	5163	8000	2837	314	288	602	52.2	47.8	no	10.2
Transcription factor with AP2 domain(s), putative	PFI1665w	603	1500	897							
Pf polyubiquitin (PfpUB)	PFL0585w	1146	2300	1154	313	7	320	97.8	2.2	no	0.6
Proliferating cell nuclear antigen 2, putative	PFL1285c	795	2350	1555							
Chaperonin, cpn60	PFL1545c	2157	4000	1843							
DNA polymerase epsilon subunit b, putative	PFL1655c	1875	3500	1625	4	82	86	4.7	95.3	no	5.0
Ubiquitin activating enzyme, putative	PFL1790w	2061	3900	1839							
DNA gyrase subunit b, putative	PFL1915w	3021	4350	1329							
Replication factor C, subunit 4	PFL2005w	1011	2600	1589	383						
Actin I	PFL2215w	1131	2500	1369	409	67	476	85.9	14.1	no	4.9

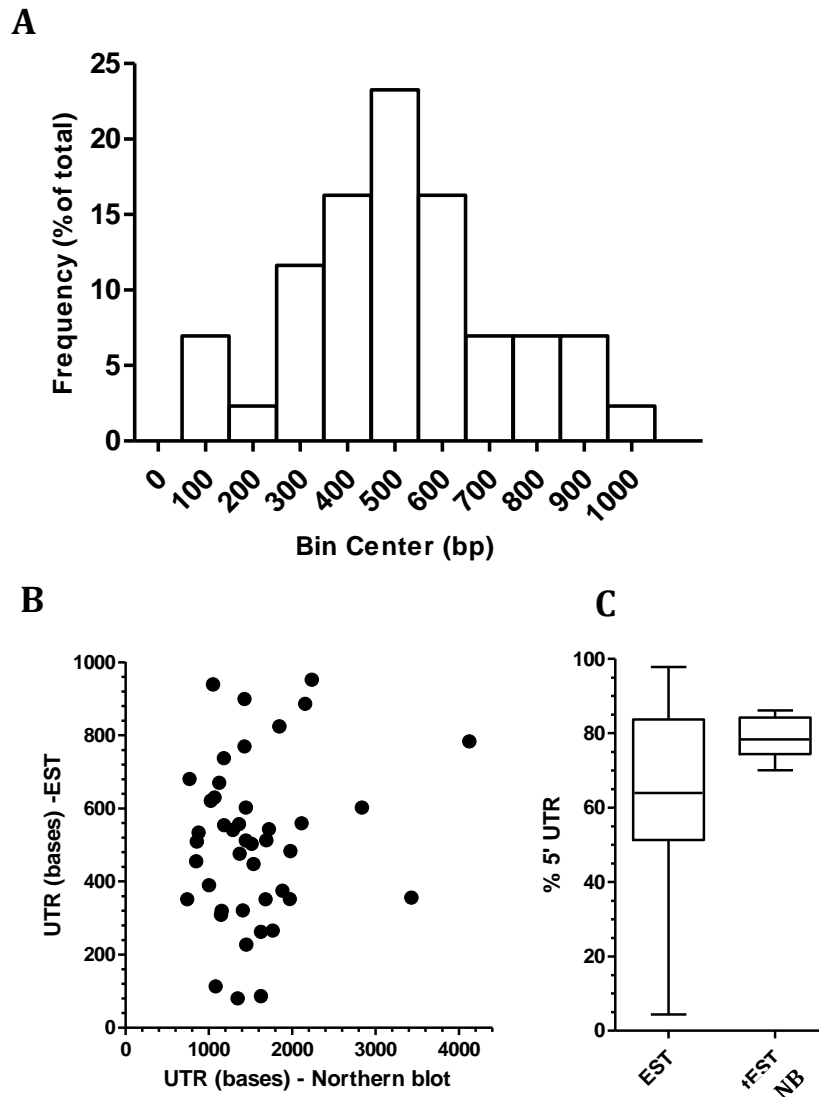


Figure 4-4 Comparative analyses of ORF for which northern blot and EST data were available

(A) Frequency distribution of predicted UTR size from EST data (% of total) for the cohort of 44 genes for which 5' and 3' EST data was available. (B) Correlation between the size of the UTR estimated from northern blot and EST datasets. (C) Box and whiskers plot of the predicted % 5' UTR calculated from the full EST dataset (EST) of 44 ORFs. The second box and whiskers plot uses the triaged EST (tEST) dataset based upon 19 ORF for which PolyA consensus transcriptional termination sequences could be identified. The graph shows the ratio of predicted 5' UTR when the tEST 3' UTR ratio was applied to the northern blot transcript size.

The apparent apportionment of the UTR within the transcript was explored finally in an attempt to better understand how transcripts are organised over genes. The available UTR

EST data (Table 4-2) indicate a median 5':3' ratio apportionment of approximately 64:36 (Fig. 4-4C – first box and whiskers plot), although the proportion of 5' UTR ranges between 4.4-70% of the predicted UTR. However, given the discrepancy between the UTR sizes provided by the northern blot and EST data, some caution must be applied to this provisional analysis of transcript apportionment. In order to better refine the dataset the 3' EST sequence data were examined more closely. *Plasmodium spp.* transcription stop sites share a consensus canonical polyadenylation site (PolyA+) motif with other eukaryotes (Cann *et al.*, 2004; Golightly *et al.*, 2000; Levitt, 1993; Ruvolo *et al.*, 1993; Wong *et al.*, 2011). Generally, these are described as 5'AAUAAA or 5'AAUAA, followed by a space of 10-30 bases, then followed again by a region relatively enhanced for G and T bases.

Of the 44 3' ESTs visual inspection of the 3' UTR and associated genomic sequence indicated that 19 share this PolyA+ motif with the remainder generally appearing to result from mis-priming from homopolymeric adenosine tracts commonly found in these AT-rich gene flanking sequences. The sizes of 3' UTR from PolyA+ ESTs ranged from 177 to 423bp in length, with those lacking this motif generally being smaller (range 5-276bp). The 3' UTR from these triaged PolyA+ ESTs (tEST) comprise an average 42.2% (range 26-95.8%) of the total EST UTR available for these 19 genes suggesting a 5':3' apportionment nearer to 58:42. However, it is likely the EST UTR size is an underestimate of the true size of the transcript – as evidenced by their consistently smaller size compared to northern blot data and potential technical issues in the data production (i.e non-processivity of reverse transcriptase and bias towards the 3' UTR from use of oligo dT in first strand synthesis). As the northern blot UTR data was also available, an alternative approach to perhaps better appreciate the relative apportionment of the predicted 3'UTR was investigated. Essentially, the northern blot UTR sizes for the 19 tEST dataset were apportioned from the high-confidence transcript termination sites allowing a predicted length of the 5' UTR to be calculated as a proportion of the total Northern blot UTR size. These data dramatically reduced the proportion of the 3'

UTR in the whole transcript (median 21.7%, range 13.9-30%) suggesting that the majority of the UTR is likely preferentially apportioned to the 5' UTR (Fig. 4-4C – second box and whiskers plot). However, as stated previously, the length of the poly-A tail was not taken into account in the *de novo* Northern blot analysis and could impact upon the apportionment ratios identified here.

4.4 DISCUSSION

This chapter attempted to explore some fundamental gaps in our knowledge regarding the transcribed region outside the ORF in *P. falciparum*, namely: the size of the UTR in a transcript, its relationship to the size of the ORF and its relative 5' and 3' apportionment. The northern blot cohort used in this analysis is relatively small representing only *c.*2% of the total genome. However, the data from the 37 northern blots produced *de novo* for this study increased our sum total of available northern blot data by some 35%. All the collated northern blot data has been submitted to PlasmoDB as user annotation data (with reference to published work). Using our relatively small dataset of 105 genes we endeavoured to correlate NB and EST data and found that whereas EST data suggested a relatively short UTR of 305 ± 182 bp the NB data indicated a much longer UTR - typically in the region of 800-1800 bases. It is worth noting that although a small dataset, the available EST data for this subset of genes indicated a 5' UTR close to that available for all genes from the Full-Malaria database (303 ± 155 bp), suggesting no inherent bias in the selected cohort (Watanabe *et al.*, 2007).

Therefore, there was a complete lack of correlation between UTR sizes predicted from northern blot and EST data - with EST data consistently shorter. One plausible interpretation for this lack of correlation could be the fact that most of the EST data was generated some 10 years ago using second generation reverse transcriptase that retained RNaseH activity (Watanabe *et al.*, 2001; Watanabe *et al.*, 2004). This combined with inherent processivity

issues over homopolymeric dA.dT sequences could impact on efficient and accurate second round PCR amplification of the oligo capped mRNA. This is perhaps evident from visual inspection of 5' ESTs that frequently terminate immediately 3' to homopolymeric poly dA.dT sequences upstream of the genes of interest. However, similar caution should also be applied to UTR sizes predicted from northern blot as these essentially represent best estimates - using electrophoretic size fractionation with a limited set of standards and correlating these data to gel migration is not going to give an exact measurement for a transcript size. In addition, transcript sizes can be particularly difficult to determine for large transcripts. However, the northern blot data did, fairly uniformly, give a long transcript size (c.800-1800bp) - much longer than the ORF and although no correlation was found between the UTR length and the length of the ORF - confirming that the UTR length is independent of ORF size, a strong correlation was observed between the Transcript and UTR size suggesting that the length of the UTR was relatively constrained.

Why *P. falciparum* UTR would be so long is matter for debate. Selective constraint upon IGR and highly conserved upstream regions, particularly for life-style associated genes (Essien *et al.*, 2008; Nygaard *et al.*, 2010) have been reported for *P. falciparum* suggesting that the content of at least some of the UTR within the transcript is likely to be well conserved. The extreme AT-bias of the IGR may influence the length of the IGR via regulatory sequence expansion within what constitutes close to a Base A and T binary code. Or, perhaps the regulatory sequences themselves do not expand but are instead embedded within other sequences which are prone to expansion; such as homopolymeric dA or dT tracts. However, although the UTR must contain all regulatory elements necessary for RNA metabolism, data presented in Chapter 7 will demonstrate that there also appears to be a significant amount of plasticity in UTR length when it comes to transcriptional and translational functionality.

Visually validating the presence of polyadenylation consensus signals for the 3' EST data gave a small (19 ORF) but robust dataset which allowed the ratio of 5' to 3' UTR to be examined. Using these ratios to interrogate the northern blot data strongly suggested that a large proportion of the UTR (perhaps as much as 70-80%) resides at the 5' end of the transcript which indicates that transcription start and stop sites could reside between 600-1300bp and 200-450bp upstream and downstream respectively. These analyses draw on the relative strengths of each methodology; northern blots provide a good estimation of total UTR length but no information regarding apportionment whereas 3' EST data that end in a consensus termination signal provides accurate information regarding the length of the 3' UTR. Interestingly, these data are in concordance with the findings presented in Chapter 3 - whereby, within this compact genome where pressure to minimize superfluous sequence must be critical, the size of the IGR, when it is required to accommodate two promoter regions is ~3 times larger than when it contains two terminator regions. From these data you could therefore hypothesize that the UTR would be likely to comply with this proportionality ratio and evidence from this limited dataset and promoter studies that have mapped transcription start and stop sites infer that this is likely the case (summary Table 1-1) (Horrocks *et al.*, 2009; Wong *et al.*, 2011).

In summary, these data provide a glimpse into the putative *P. falciparum* transcriptional landscape through systematic comparison of available data. These findings have implications for *in silico* searches - most of which employ a 1000bp search window. If the transcription start site is truly this far upstream, for many genes, a 1000bp search window may not provide adequate coverage for the identification of all putative upstream regulatory elements. In addition, if the UTR are truly this long - *c.*800-1800bp - this would suggest that 40-90% of all IGR may be a part of at least one transcript and hints that transcriptional unit overlap may occur. A modelling approach to investigate how the UTR could be accommodated within the available intergenic space is the topic of Chapter 5.

CHAPTER 5 MODELLING SPATIAL ORGANIZATION OF TRANSCRIPTS OVER INTERGENIC SPACE IN *P. FALCIPARUM*

5.1 INTRODUCTION

The two previous chapters have:

- (i) Explored the size of intergenic regions (IGRs) in *P. falciparum* and shown that the size of these regions relates to the nature of the transcriptional activity that occurs over them. It has been demonstrated that IGRs flanked by ORFs in a Divergent:Tandem:Convergent configuration (termed A:B:C respectively) demonstrate a 3:2:1 spacing ratio for moderately compact genomes.
- (ii) Analysis of available northern blot and EST data suggests that UTR are long (in the region of 800-1800bp) and that they are likely apportioned in an approximate 78:22 upstream to downstream ratio over the ORF, respectively.

Little is known about the nature of transcriptional units over these IGR. As mentioned previously, available RNASeq data have proved extremely difficult to allocate to these extremely AT rich and repetitive sequence regions and although longer sequence reads and improved amplification/bioinformatic strategies are being developed, we currently have only a limited amount of *in vitro* data available (Otto *et al.*, 2010; Sorber *et al.*, 2011; Ponts *et al.*, 2012).

In 1992, using nuclear run-on, Lanzer *et al.*, described transcript termination and re-initiation over an IGR residing between the *3.8* and *GBP130* ORF. These data indicated that two discrete non-overlapping transcripts were produced from within this IGR – confirming transcription in *P. falciparum* was monocistronic in nature (Lanzer *et al.*, 1992b). Interestingly, the UTR size reported in this study for promoter and terminator regions are in accordance with our

predictions for relative transcript apportionment. However, nuclear run-on can be technically challenging; it can be difficult to label nascent RNA and normalise the hybridisation for such AT-rich sequence which has limited the widespread adoption of this approach to study transcriptional arrangements. Although many other functional studies have been undertaken most tend to consider promoter (transcriptional activity toward the reporter gene) or terminator function in isolation (reviewed in Horrocks *et al.*, 2009) and do not take into account the fact that IGR act as transcriptional regions for both flanking genes.

In 2002, Kyes *et al.*, detected, via northern blot probed with dsDNA over a life-cycle time-course, the presence of two transcripts for MSP2; a 2kb and a 4.4kb transcript. The 2kb transcript was highly expressed, as expected, during the later life-cycle stage (30-38hpi). The 4.4kb transcript, however, was expressed at ring and early trophozoite stages. Further investigation identified that this transcript initiated from the adjacent tail-to-tail gene - predicted to be adenylosuccinate lyase (ASL). In total 3 transcripts were detected from these two genes; (i) a 2kb MSP2 transcript expressed 30-38hpi, (ii) A 3kb ASL transcript expressed 3-30hpi and (iii) a 4.4kb sense ASL/antisense MSP2 'readthrough' transcript expressed at 30-38hpi (Kyes *et al.*, 2002). Therefore, although transcription in *P. falciparum* generally appears to be monocistronic, the presence of a sense/antisense 'readthrough' transcript suggests that this may not always be the case.

A myriad of non-coding and antisense transcripts (ncRNA) have also subsequently been detected as being expressed during the *P. falciparum* IE cycle (Gunasekera *et al.*, 2004; Upadhyay *et al.*, 2005; Militello *et al.*, 2005; Chakrabarti *et al.*, 2007; Li *et al.*, 2008; Mourier *et al.*, 2008). In the apparent absence of the major Dicer and Argonaute proteins - suggesting that miRNA are not present in this parasite (Ullu *et al.*, 2004; Cerutti and Casas-Mollano, 2006; Militello *et al.*, 2008), the true variety and class functionality of ncRNA remains to be fully elucidated. However, a wide range of ncRNA do appear to be transcribed and specific

roles, particularly in telomere gene maintenance and gene virulence regulation (Broadbent *et al.*, 2011; Sierra-Miranda *et al.*, 2012), putative roles in the function and maintenance of centromeric chromatin (Li *et al.*, 2008) and control of RNA expression (Sims *et al.*, 2009) are beginning to emerge.

Comprehensive transcriptomic datasets are available for the *P. falciparum* IE cycle (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Otto *et al.*, 2010) demonstrating a potentially 'hard-wired' cascade of gene expression with single peaks of mRNA abundance. However, how these mRNAs correlate spatially with the transcriptional landscape over the IGR remains unclear; for example, whether transcription and termination of transcripts within all IGR similarly comprise discrete non-overlapping entities, as shown for *gbp130* and *3.8*, is unknown. Our data indicates that long UTR are generated over the relatively small IGR of the compact *P. falciparum* genome - strongly suggesting that transcript overlap may occur. It is estimated that over 10% of predominantly lineage-specific genes in humans 'overlap' (Makalowska *et al.*, 2005; Osato *et al.*, 2007; Sanna *et al.*, 2008) and that these gene 'overlap' events can fall into many classes such as: Head to Head (promoter overlap), Tail to Tail (terminator overlap), Nested (same or different strand) or same-strand gene overlap events (Ho *et al.*, 2012). These data suggest that transcriptional overlap may in fact not be an unusual phenomenon.

In order to better understand the nature of the transcriptional landscape it was decided to attempt to model spatial organization of the UTR over the IGR. We use a simple modelling approach to explore two key concepts relating to the nature of transcription over an IGR in *P. falciparum* asking;

- i) Can we provide supporting evidence, from a global analysis of transcript apportionment, for the preferential organisation of the UTR upstream of the ORF.

- ii) Can we determine whether it is likely that transcriptional units are discrete, non-overlapping units, or, do they likely overlap given the apparent large size of UTR in the relatively compact *P. falciparum* genome.

5.2 MODELLING APPROACH

Two criteria were assessed:

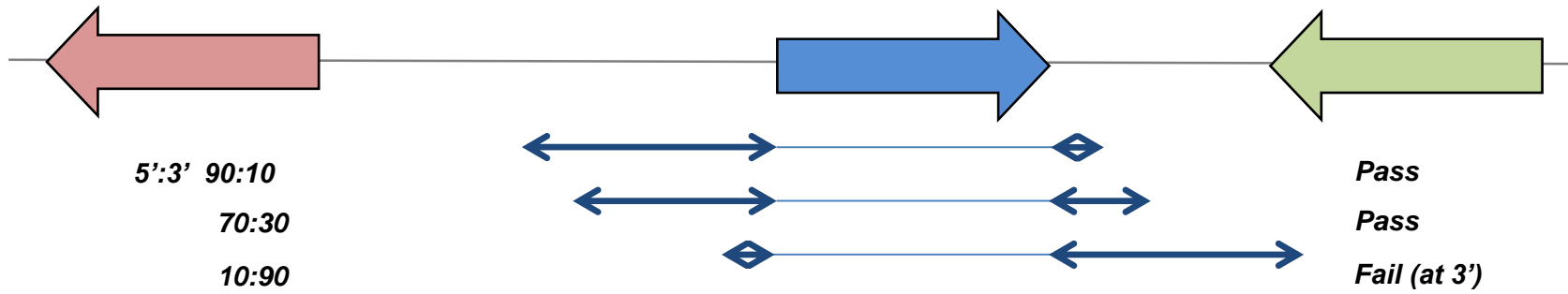
1. Scenario A

Assumes the transcript can only start/stop within an IGR - hence no overlap of adjacent ORFs is tolerated - this is considered a fail. Scenario A takes each consecutive gene triplet from each chromosome in isolation using each ORF sequentially as the central gene of the triplet (with the exception of the first and last gene on each chromosome). It then apportions the UTR at 1% increments 5' to 3' over the available intergenic space on either side of the central ORF. Therefore, it explores the fit and preferential apportionment of the UTR over each gene only (Fig. 5-1).

2. Scenario B

This is a more constrained model. It still assumes, as above, the transcript can only start/stop within an IGR - hence no overlap of adjacent ORFs is tolerated this is considered a fail. However, Scenario B also records a fail when a transcript overlap event occurs from either of the similarly apportioned UTR from the two flanking gene triplets. So this model explores the fit and preferential apportionment of the UTR over all genes in the genome and also asks whether the IGR is able to be utilized mutually exclusively at each 5' to 3' ratio by the two flanking ORF respectively (Fig. 5-2).

Scenario A



Scenario B

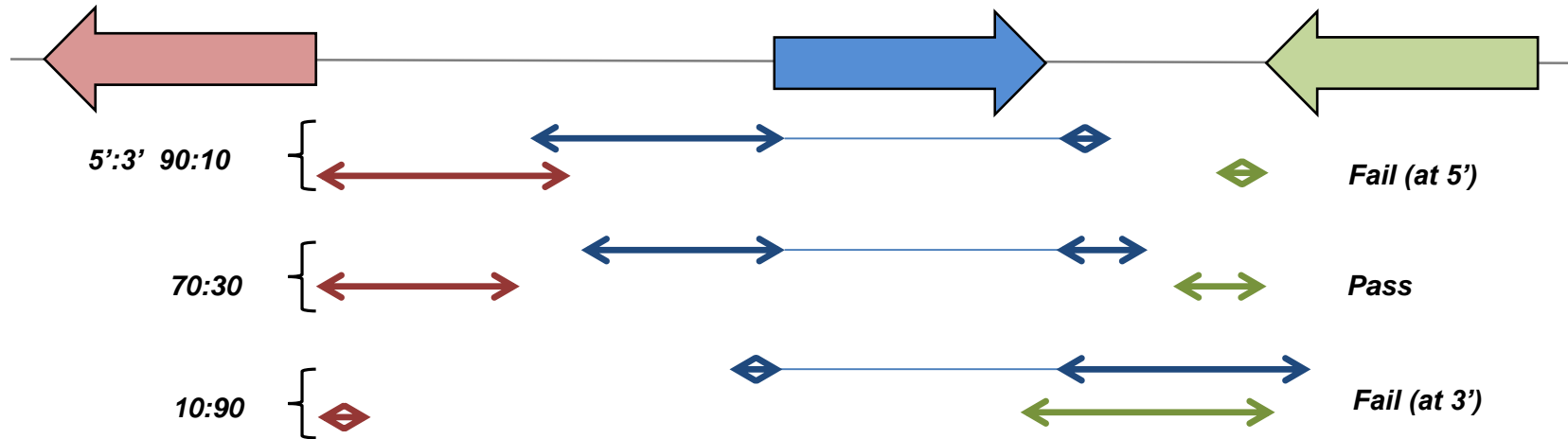


Figure 5-1 Scenario A

(previous page) Takes the IGR from each gene triplet and apportsions the specified UTR size in 1% increments over the available IGR. A pass (0) is recorded if there is no ORF overlap and a fail (1) recorded if there is. A pass/fail is recorded for each 5' to 3' apportionment ratio.

Figure 5-2 Scenario B

(previous page) Takes the IGR from 3 gene triplets (only the central gene triplet is shown) and apportsions the specified UTR size in 1% increments over the available IGR. A pass (0) is recorded if there is no ORF or adjacent (similarly apportioned) UTR overlap and a fail (1) recorded if there either of these scenarios occur. A pass/fail is recorded for each 5' to 3' apportionment ratio.

Firstly, it was necessary to capture the information relating to the size and type of UTR available for each of the *P. falciparum* genes within its genome and this information was contained within the general format file (.gff) file in the form of ORF co-ordinates and gene orientation (downloaded from PlasmoDB). Two *de novo* PERL scripts; intergenic.dist.2CLASH.pl and intergenic.dist.2CLASH.incl.flanking.pl were then designed and written to extract the relevant information from the .gff file and apportion the UTR (UTR size specified by the user in bp) over the available space within the IGR (Fig. 2-4 Materials and Methods shows the overall workflow). Fig. 5-3A shows a schematic of the information required for extraction.

Essentially these two PERL scripts took the size of the IGRs for each gene triplet from the parsed .gff file and apportioned the (user specified) UTR size at 1% increments from 5' to 3' over the available intergenic (IG) space. For each IG space ratio a binary pass/fail was recorded. If the UTR fitted into the available space at the specified ratio it scored a 0; if it did not fit into the space because of an ORF 'clash' or overlap (scenario A and B) and/or an adjacent UTR overlap (scenario B only) it scored a 1. Fig. 5-3B demonstrates the decision making process behind the two Perl scripts. Fig. 5-4A shows a fractioned portion of the script output demonstrating that: the gene triplets are named, the intergenic co-ordinates and DNA strand from which they originate are recorded, as are the orientation of the flanking genes

and the total IG region in bps. It should be noted that the orientation data is still in the H2H, H2T/T2H and T2T format which correlates to: A-type (Divergent), B-type (Tandem) and C-type (Convergent) respectively. Fig. 5-4B (also part of the script output) demonstrates the binary pass/fail system (UTR size apportioned 100% 5'/0% 3' to 0% 5'/100% 3') using, in this case, a 1.4kb UTR size. The % cumulative fail rate was then calculated (the sum of the binary output) and this was graphed against 5% occupancy.

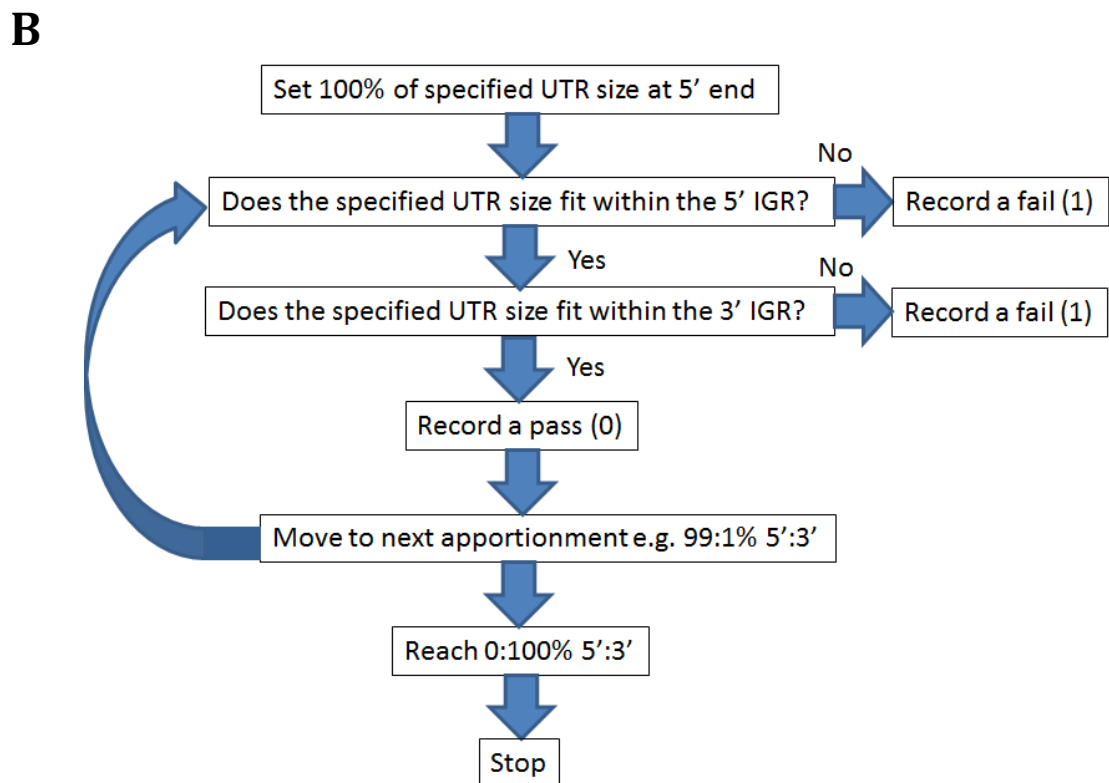
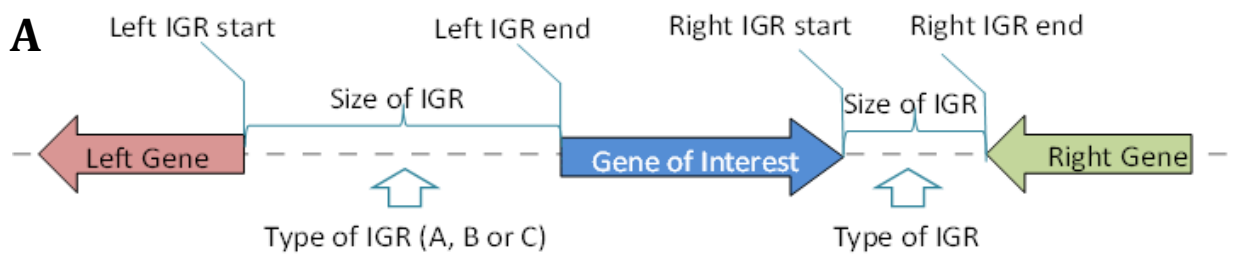


Figure 5-3 Input data for and methodology and decision making processes of intergenic.dist.2CLASH.pl and intergenic.dist.2CLASH.incl.flanking.pl

(previous page) (A) demonstrates a graphical representation of the input data required arrayed over the gene triplet. (B) demonstrates the methodology and decision making processes of intergenic.dist.2CLASH.pl and intergenic.dist.2CLASH.incl.flanking.pl. The two algorithms work in incremental 1% 5':3' ratios. Initially 100% of the user specified UTR is allocated to the 5' IG region and through an assessment of UTR fit (of size x) within available intergenic space they assign a pass (0) or fail (1) to each 5':3' apportionment ratio.

A

Left Gene	Left Gene (Start_End_strand)	inter_type(Left-Current)	Current Gene	Current Gene (Start_End_strand)	inter_type(Current-Right)	Right Gene	Right Gene (Start_End_strand)	Upstream dist(bp)	Downstream dist(bp)
PF13_0194	1411298 1411906 -	H2T	MAL13P1.174	1413300 1414145 -	H2T	MAL13P1.175	1415076 1415216 -	931	1394
MAL13P1.184	1491124 1494504 -	H2H	MAL13P1.185	1495455 1497436 +	T2H	MAL13P1.186	1498073 1501690 +	951	637
MAL7_rRNA_Thr2	1374387 1374458 -	H2H	MAL7P1.170	1376682 1377743 +	T2H	MAL7P1.171	1380751 1387189 +	2224	3008
MAL7P1.231	98671 99734 +	T2H	MAL7P1.228	106224 108526 +	T2T	MAL7P1.229	110392 115696 -	6490	1866
MAL7P1.230	92135 93136 -	H2H	MAL7P1.231	98671 99734 +	T2H	MAL7P1.228	106224 108526 +	5535	6490
PF07_0034	454668 455621 -	H2H	MAL7P1.27	458600 461695 +	T2H	PF07_0035	463593 467339 +	2979	1898
PF07_0063	749899 750258 +	T2T	PF07_0064	750976 752280 -	H2H	PF07_0065	754615 756285 +	2335	718
PF07_0064	750976 752280 -	H2H	PF07_0065	754615 756285 +	T2H	PF07_0066	758583 761960 +	2335	2298
MAL8P1.40	1023701 1024918 -	H2H	PF08_0034	1029824 1034586 +	T2H	MAL8P1.38	1034637 1037500 +	4906	51
MAL8P1.153	152929 160789 -	H2H	PF08_0131	166536 167198 +	T2T	PF08_0130	168189 171554 -	5747	991
PF10_0015	68690 68962 -	H2T	PF10_0016	70548 70820 -	H2T	PF10_0017	72935 73918 -	2115	1586
PF10_0083	357345 358463 +	T2T	PF10_0084	360625 362480 -	H2T	PF10_0085	365737 367146 -	3257	2162
PF10_0142	576578 580477 +	T2T	PF10_0143	581353 589089 -	H2H	PF10_0144	595307 596221 +	6218	876
PF10_0152	624047 625894 -	H2T	PF10_0153	627043 629039 -	H2T	PF10_0153a	629908 632060 -	869	1149
PF10_0153a	629908 632060 -	H2H	PF10_0154	633681 635285 +	T2H	PF10_0155	637137 639010 +	1621	1852
PF10_0158	646053 647735 +	T2T	PF10_0159	650897 653566 -	H2T	PF10_0160	656708 659026 -	3142	3162
PF10_0159	650897 653566 -	H2T	PF10_0160	656708 659026 -	H2T	PF10_0161	661741 666358 -	2715	3142
PF10_0164	682181 682507 -	H2H	PF10_0165	685577 688861 +	T2T	PF10_0166	689887 690819 -	3070	1026
PF10_0191	804327 807581 -	H2T	PF10_0193	808617 808991 -	H2T	PF10_0194	811105 813483 -	2114	1036

B

100	99	98	97	96	95	94	93	92	8	7	6	5	4	3	2	1	0
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

5.3 RESULTS

5.3.1 TESTING OF SCENARIO A

In order to be sure that our model was working correctly it was tested against the 105 ORF for which we had northern blot data available – hence a UTR size. It is worthy of note that the northern blot data originated from a diverse range of gene types, gene structure and from varied life-cycle temporal expression profiles. The IGR data for this gene cohort was run through `intergenic.dist.2CLASH.pl` using the predicted 1.45kb UTR size (the y-intercept from the linear regression) and Fig. 5-5A shows the plot of ‘real’ v. modelled data. It was important to see whether the whole genomic dataset could be encapsulated within a single value and Fig. 5-5A demonstrated very good correlation between the ‘real’ northern blot data and the modelled data. The 5’/3’ apportionment ratio (the lowest point on the graph represents the optimal 5’ apportionment) for the modelled data and predicts a preferential apportionment close to the 78:22% 5’/3’ as also estimated from the triaged EST/northern blot data. In order to ensure this apportionment ratio was not an artefact of using the predicted UTR size, three further UTR sizes were also analysed and plotted for this dataset; 1.2kb 1.4kb and 1.6kb and the data are shown in Fig. 5-5B. These data demonstrated that, for the 105 gene dataset, although as expected the ‘fit’ of the data was not as good, the 5’ to 3’ apportionment ratio did not alter significantly.

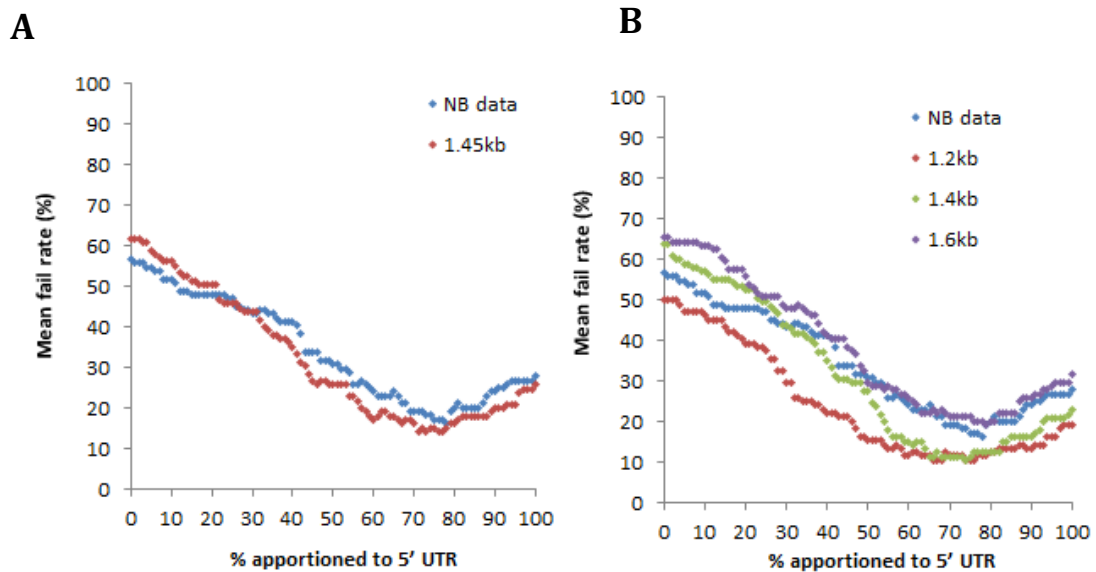


Figure 5-5 Intergenic.dist.2CLASH.pl output data for the 105 northern blot gene cohort

The same cohort of 105 genes for which northern blot was available (see Table 4-1) fitted with a 1.45kb UTR. These data are plotted as the mean fail rate against the % of the UTR apportioned to the 5' end of the transcript. Hence the nexus in the graph represents the ratio of best fit. In figure 5B a modelled 1.2, 1.4 and 1.6kb transcript are used for comparative purposes.

5.3.2 SCENARIO A

With the reassurance that our Perl script was working correctly and accurately predicting anticipated results, the whole *P. falciparum* IGR genomic dataset was interrogated. Using intergenic.dist.2CLASH.pl (Scenario A) the UTR fit over the IGRs for the whole *P. falciparum* genome were modelled using different predicted UTR sizes; 600bp to 1800bp in 200bp increments and the output data is depicted in Fig. 5-6.

600-1800bp UTR sizes were used as this captured 80-90% of the UTR sizes observed within the northern blot dataset and the 200bp increments allowed the shape of the graph to be explored without crowding the graph. The outcome was perhaps not unexpected in that the larger the predicted UTR size the higher mean % fail rate - with optimum apportionment fail

rates ranging from 10.2 to 47.8% (600 and 1800 bps respectively). What was interesting was that even with a 600bp UTR, close to the minimum UTR size - as determined from the northern blot data, the fail rate at optimum apportionment never reaches 0%. However, previous data have demonstrated that median IGR sizes vary depending upon the chromosomal compartment in which they reside (Chapter 3). It is also possible that the median length of the IGR may differ according to other criteria - such as ORF functionality.

The fact that Figure 6 shows a series of similarly shaped curves suggests that although the 'fail rate' increases with an increase in predicted UTR size, the optimal 5'/3' UTR apportionment stays fairly consistent at around 80:20% 5'/3' - although the optimum apportionment range does increase as the UTR length decreases.

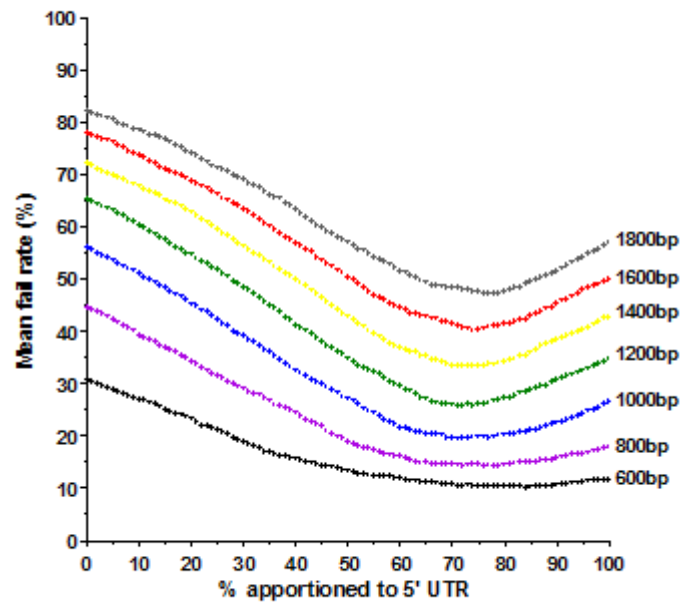


Figure 5-6 Plot of mean % fail rate against % UTR apportioned to the 5' for the whole *P. falciparum* genome using predicted UTR sizes of 600 to 1800bp (200bp increments) using Scenario A.

By granulating the data it was hoped that this would allow us to glean more information regarding UTR fit and apportionment. The data were sorted into 4 groups dependent upon the orientation of the genes that flanked the IGR. Group 1 represents a B/A or a reverse A/B situation, Group 2 represents a C/A or a reverse A/C situation, Group 3 represents a B/B or a reverse B/B situation and Group 4 represents a C/B or a reverse B/C situation and these groupings are depicted graphically in Fig. 5-7.

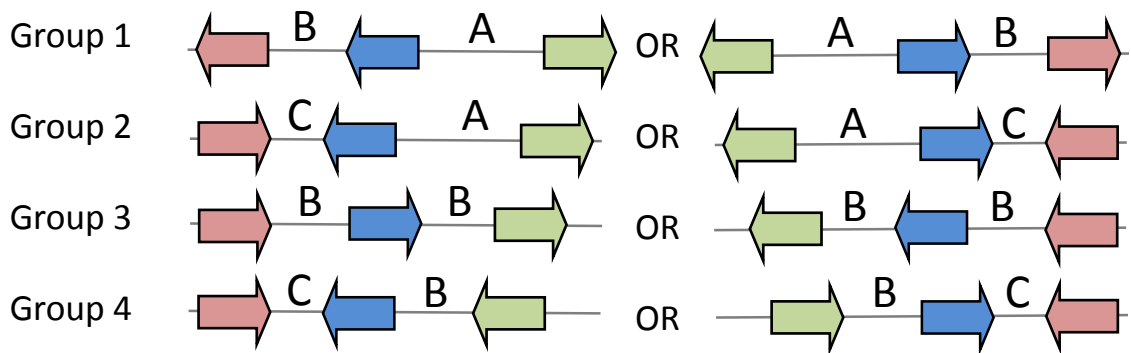


Figure 5-7 Visual representations depicting the median IGR size dependent upon the flanking gene orientation, and the orientation of the three genes within each triplet (forward and reverse) that are contained within each group.

Figure 5-8 shows the output of such a grouping; again the % 5' apportioned UTR is plotted against the mean % fail rate. Previous work has demonstrated a clear 3:2:1 median genomic spacing ratio for IGRs flanked by ORFs in a A:B:C orientation respectively. What this granulation of the data clearly shows is how the composite whole genome figure is reached. In Group 1, an A/B (or B/A) type IGR combination provide a maximal and an intermediate size IGR and therefore the lowest fail rates. Whereas in Group 3, a B/B type IGR combination provide two intermediate sized IG spaces and therefore intermediate fail rates. However, when an attempt is made to fit 1.4kb UTR in its entirety into a smaller C-type IGR up to a 90% fail rate is observed (Groups 2 and 4 respectively). Perhaps not unexpectedly, the optimum 3' to 5' apportionment (denoted by the nadir of the line) varies for each group depending upon

the flanking IGR types. However, all IGR type combinations still appear to express an optimum 5' apportionment preference.

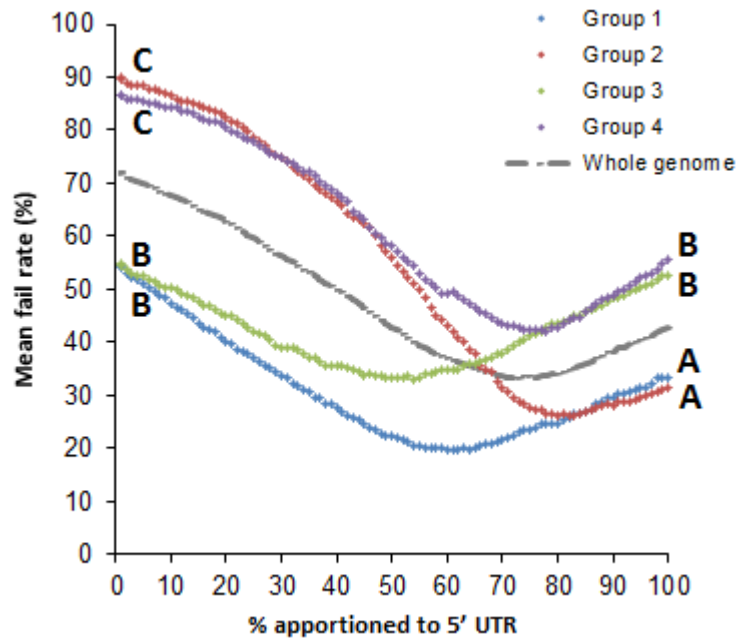


Figure 5-8 Granulation of intergenic.dist.2CLASH.pl output data

Each of the groups describe an orientation pair (forward and reverse) for the genes flanking the two IGRs over the gene triplet. Group 1 represents a H2T/H2H or a reverse H2H/T2H situation, Group 2 represents T2T/H2H or a reverse H2H/T2T situation, Group 3 represents a T2H/T2H or a reverse H2T/H2T situation and Group 4 represents a T2T/H2T or a reverse T2H/T2T situation. The plot is a granulation of the data output for Scenario A using a predicted UTR size of 1.4kb. The whole genome dataset using a predicted UTR size of 1.4kb is also shown for comparison.

5.3.3 SCENARIO B

As previously outlined Scenario A was the least constrained of the two scenarios and looked at each gene triplet in isolation – it only considered an ORF overlap a fail. Scenario B was much more stringent and took into account the UTR of adjacent gene triplets; failing not only if there was an ORF overlap but also if there was a UTR overlap. Intergenic.dist.2CLASH.incl.flanking.pl (Scenario B) was applied to the same *P. falciparum* IGR

genomic dataset with the same predicted UTR sizes (600-1800bp) in 200bp increments and these data are shown in Fig. 5-9.

What was immediately observed was the much larger mean % fail rates observed using this scenario - regardless of the size of the UTR. For Scenario B the minimum fail rates ranged between 23.2 and 81.8% (600 and 1800bps respectively). These data suggest that either the size of the UTR is considerably smaller than previous data has indicated - the nadir of even a 600bp UTR never reaches 0, or that there is substantial UTR overlap occurring within the IGR. Interestingly, the optimum % 5' UTR apportionment appears not to vary profoundly although it does appear to gravitate from around 70% for a 600bp predicted UTR to around 90% for an 1800bp predicted UTR.

These data again strongly support previous UTR apportionment findings indicating a ~ 80:20% 5':3' UTR apportionment preference. They also suggest that sequence within an IGR may not be mutually exclusive to the UTR of one of the flanking genes only.

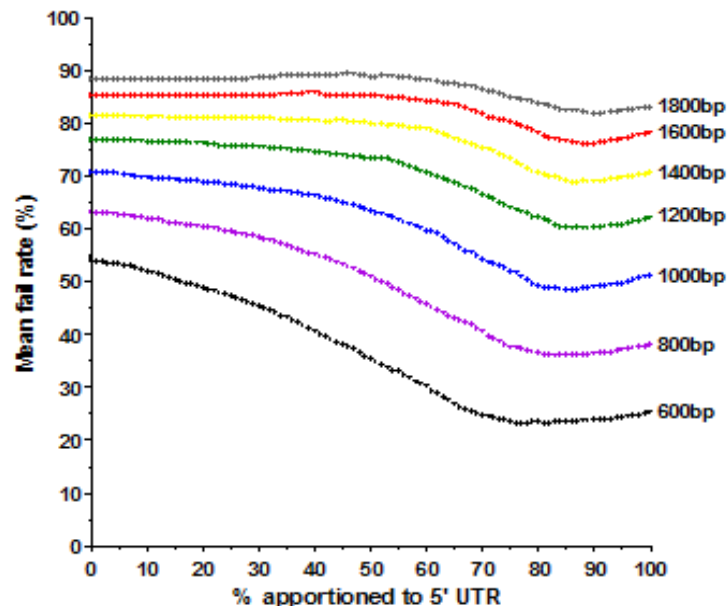


Figure 5-9 Plot of mean % fail rate against % UTR apporportioned to the 5' for the whole *P. falciparum* genome using predicted UTR sizes of 600 to 1800bp in 200bp increments using Intergenic.dist.2CLASH.incl.flanking.pl (Scenario B).

5.4 GENES IN A HEAD TO HEAD ORIENTATION WITH THE SAME TEMPORAL PROFILE HAVE SMALLER IGRS

The data modelling, using scenario B, was suggestive of transcriptional overlap events occurring over IGR. This raised another question: significant transcriptional unit overlap would then open up the potential for steric hindrance as two RNA Pol II complexes attempt to use opposing strands of DNA simultaneously. Previously published expression data (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Otto *et al.*, 2010) has clearly demonstrated the IE cascade of gene expression, with each transcript exhibiting one temporal peak of amplitude per life-cycle – could it be that co-transcription over IGRs is temporally arranged to minimise conflict?

The availability of such comprehensive IE stage-specific datasets for *P. falciparum* (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Jurgelenaite *et al.*, 2009; Otto *et al.*, 2010) enabled the exploration of the incidence of co-spatial co-temporal expression in order to ascertain how frequently such events occur and whether such events leave a ‘footprint’ on the size of the IGR. Although other datasets were more comprehensive, attempts at dividing these data into small temporal groups yielded data subsets too small for meaningful analyses. Therefore, the Jurgelenaite *et al.*, dataset was selected for temporal division as this group had clustered the genes into 4 life-cycle expression stages; i) early ring, ii) late ring and early trophozoite, iii) trophozoite and early schizont and iv) schizont and v) a ‘constitutively IE expressed’ bin/cluster (Jurgelenaite *et al.*, 2009).

Essentially the available A, B, C IGR data were overlaid with an additional layer of information – a time ‘window’ of transcription. This was expressed numerically according to the life-cycle stage whereby 1 = early ring (144 ORF), 2 = late ring/early trophozoite (1033 ORF), 3 = trophozoite/early schizont (985 ORF), 4 = schizont (329 ORF) and 5 = ‘constitutive’ IE expression (1344) (Jurgelenaite *et al.*, 2009). This numerical ‘labelling’ of adjacent gene pairs

in A:B:C IGR datasets enabled the data to be filtered into co-located and co-expressed sub-groups.

The whole genomic IGR dataset gave an A:B:C ratio of 1479:2626:1483 (Table 5-1). When cluster 5, 'constitutively IE expressed' ORF were included, this gave an A:B:C ratio of 646:885:621. However, taking only cluster 1, 2, 3 and 4 life-cycle expression data - co-located and truly co-expressed (same temporal window) - the A:B:C ratio reduces to 202:237:129 (Table 5-1) - a mere 13.65% (A), 8.59% (B) and 8.61% (C) of the total genomic IGR. These data suggest that gene orientation and transcriptional timing may in fact be structured to minimise transcriptional conflict.

Table 5-1 Comparison of intergenic region size for whole *P. falciparum* genome and co-located co-transcribed genes only

Domain	IGR Type	n=	Ratio of IGR		
			types	Median Size	
			Ratio of median size		
All genome	A	1479	1.00	1938	2.86
	B	2626	1.77	1385	2.05
	C	1483	1.00	677	1.00
Co-transcribed and collocated (c/c)	A	202	1.57	1539	2.18
	B	237	1.84	1428	2.03
	C	129	1.00	705	1.00

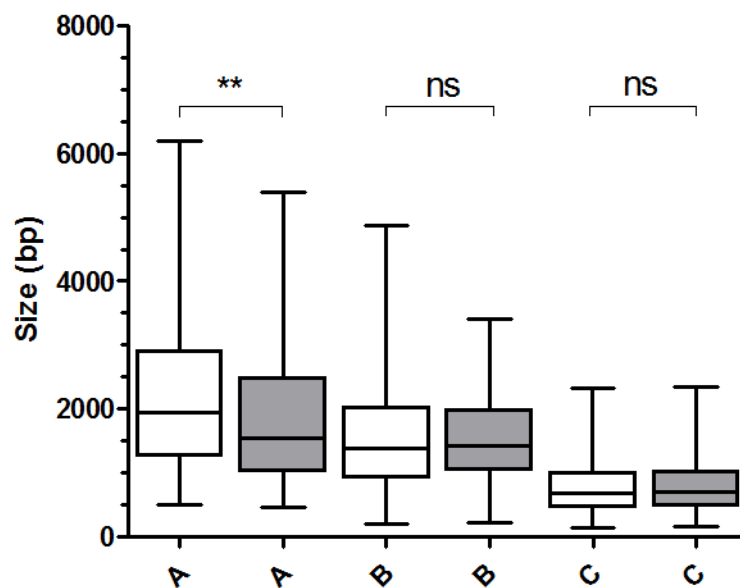


Figure 5-10 Analysis of temporal co-transcription using *P. falciparum* intergenic region size

For each IGR type, the result of an ANOVA with Dunn's multiple comparison post-test (** P < 0.01, ns, not significant) is shown. Clear boxes represent the size distribution of all IGR. Grey boxes represent the IGR sizes over which two transcripts occur within the same temporal IE window. Outliers beyond 2.5-97.5% of the data represented by the whiskers are not shown.

The size of the A:B:C co-temporal co-expressed IGR dataset were then analysed and compared with the size of the whole genomic IGR dataset (as per Chapter 3) and the results are shown in Fig. 5-10. The A>B>C relationship still existed with median A:B:C sizes of 1539,

1428 and 705bp respectively, but, whereas there was no significant difference between the size of the IGR between the whole genomic or co-located co-transcribed datasets for type B and C IGR, type A co-transcribed co-located IGR were significantly smaller than their genomic family as a whole. This was an interesting outcome – as you would perhaps expect a co-transcribed co-located A-type IGR, having to accommodate two promoter regions, to be larger not smaller.

5.5 DISCUSSION

Using a modelling approach we have shown that, using a fixed length UTR input, we can optimally apportion UTRs over IGR for the whole *P. falciparum* genome. Our data demonstrate that whereas minimum fail rates were relatively low for the less stringent Scenario A (10.2 to 47.8% 600-1800bps UTR respectively), they were considerably higher when using the more stringent Scenario B (23.2 to 81.8% 600-1800bps UTR respectively). These data combined suggest that substantial spatial overlap of UTR may occur within *P. falciparum* IGR. It is also plausible, from the Scenario A fail rates, that transcriptional start sites may not be located solely within the IGR.

The methodology used for these analyses records a pass/fail binary output to define discrete parameters. Variably sized and apportioned UTR (although length and apportionment variance is undoubtedly a more likely scenario) have not been considered as this would have required a considerably more complex algorithm; taking into account randomization. The other issue with such an approach would have been the creation and analysis of a huge amount of data – perhaps a PhD in itself. Therefore, essentially this is a reductionist model and some caution needs to be applied before the extrapolation of these findings to real-life. However, transcriptional overlap is not a unique phenomenon. It is common in microbial genomes (Johnson and Chisholm, 2004) and present in eukaryotic organisms as diverse as yeast (Hermsen *et al.*, 2008) and humans (Ohinata *et al.*, 2002; Edgar, 2003; Zhou and

Blumberg, 2003; Veeramachaneni *et al.*, 2004; Ho *et al.*, 2012) and the full complexity of the non-protein coding transcriptional landscape; IGR space apportionment and usage, IGR sequence conservation and the identification of and roles for ncRNA are only just beginning to be explored (Militello *et al.*, 2008; Chakrabarti *et al.*, 2007; Mourier *et al.*, 2008; Essien *et al.*, 2008; Hermsen *et al.*, 2008; Nygaard *et al.*, 2010; Ho *et al.*, 2012).

Hermsen *et al.*, identified a similar situation in yeast. This group modelled upstream coding regions (UCRs) and downstream coding regions (DCRs) in *S. cerevisiae* using two constraints; (i) ORFs do not overlap and (ii) although UCRs and DCRs may overlap with each other or ORFs – this event was allocated a ‘less probable’ factor of q . What this group identified was that although their model fitted IGR from convergent genes (C-type IGR) and tandem genes (B-type IGR) it did not fit one of two identified populations of divergent genes (A-type IGR). This led to the assumption that this population probably represented a subset of bi-directional promoters (Hermsen *et al.*, 2008). Interestingly, this group also declared that the assumption that UCRs and DCRs are of fixed length was a limitation in their results.

Ho and colleagues, as mentioned previously, have looked at different configurations for gene-overlap events in humans and this group identified that the highest co-expression levels emanated from the Promotor Overlap group - probably as a result of the sharing of local chromatin status and the lowest co-expression level emanated from the Tail-to-Tail (T2T) Overlap group – most of which share polyadenylation signal overlap. In the T2T situation it was surmised that sense/antisense overlap of polyadenylation motifs may cause transcriptional interference. Interestingly, co-regulation increases if the overlapping region extends into the first exon/intron (Head to Head Overlap group) and whereas most of the genes in a H2H or (Type A) configuration were lineage specific most of the genes within the T2T or (Type C) configuration were non-human specific genes (Ho *et al.*, 2012).

These data could perhaps go some way to explaining why we see a significantly shorter length with the co-spatial co-temporal A-type IGR but not a B- or C-type IGR. Further analysis on the component genes of the A-type group would be interesting to identify whether this group is comprised of predominantly lineage-specific genes also. It is possible that these 202 co-spatially arranged co-temporally expressed type-A IGR could represent the start of a short-list of *P. falciparum* bidirectional promoters. Bi-directional promoters have been identified and characterized in other organisms from humans (Zanotto *et al.*, 2009) to *D. discoideum* (Hirose *et al.*, 2006) therefore, although not common, are once again likely to be a universal phenomenon.

Interestingly, studies in yeast have identified the presence of promoter divergent cryptic unstable RNA Pol II transcripts (CUTs) which are normally targeted for degradation (Neil *et al.*, 2009). In 2009 Neil and colleagues published data taken from a study of *S. cerevisiae* glycolysis genes. The combination of these data led this group to surmise that eukaryotic promoters may actually be intrinsically bi-directional - with some of the resultant transcripts possibly playing a role in gene regulation (Neil *et al.*, 2009). If Eukaryotic promoters are 'intrinsically bi-directional' it is plausible that the transcripts that are produced during different temporal windows from A-type IGR could partly explain the large fail rates observed in Scenario B as these IGR may contain considerable areas of 'overlap'. It is also plausible that the presence of low-level divergent transcripts for the flanking ORF may play a role in the regulation or silencing of this (adjacent flanking) ORF during this temporal window.

In summary co-spatial co-temporal expression does occur in *P. falciparum* – however, it does not appear to be a frequent event (only ~10% of total genomic IGR from this study shared the same unique temporal window of transcription and a co-spatial flanking ORF arrangement). The data used for these analyses also utilized broad temporal life-cycle stage windows of 8-12hrs – therefore it is also plausible, given the highly co-ordinated cascade of gene

transcription that this parasite exhibits, that co-spatial co-temporal transcription may occur even less frequently. The UTR modelling data is strongly suggestive of the occurrence of UTR overlap events. But these data do not support a hypothesis that co-spatial *P. falciparum* ORFs are transcribed simultaneously. Instead, they suggest that 'shared' promoter or terminator sequences may exist within the IGR servicing different ORF during different temporal windows. Overlapping genes are also common in bacteria, mitochondria and viruses (Johnson and Chisholm, 2004) and are present at over 10% within the human genome (Sanna *et al.*, 2008). These may take the form of same- or different-strand nested ORF overlap events or promoter- or terminator-overlap events potentially adding another layer of complexity (Sanna *et al.*, 2008). Of note perhaps, is that the presence of nested promoters in *P. falciparum* could perhaps explain the detection of the Tata Binding Protein (PFTBP) within coding sequence reported by Ruvalcaba-Salazar *et al.* (Ruvalcaba-Salazar *et al.*, 2005).

These data provide an insight into the transcriptional landscape of the non-coding *P. falciparum* genome. They support previous findings indicating that 70-80% of the UTR is apportioned at the 5' end of the transcript and suggest the occurrence of extensive UTR transcriptional overlap. Elaboration and extension of this approach to incorporate UTR length and apportionment randomization and the necessary mega-data analysis, perhaps combined with data from ever-improving RNA-Seq datasets, could produce some very interesting results and perhaps bring us a little closer to understanding the non-protein coding *P. falciparum* transcriptome.

CHAPTER 6 ORGANIZATION OF HOMOPOLYMERIC dA.dT TRACTS IN *P. FALCIPARUM* AND OTHER APICOMPLEXAN PARASITES

6.1 INTRODUCTION

Homopolymeric poly dA.dT tracts are found at prolific frequency within the *P. falciparum* genome, particularly within the AT-rich intergenic regions (Dechering *et al.*, 1998; Zhou *et al.*, 2004). Whether these poly dA.dT tracts are simply a facet of the extreme AT richness, whether there is an organisation to such tracts or whether they impart functionality in *P. falciparum* remains, as yet, unclear. This chapter aims to qualify some of these points by providing a comparative spatial analysis of homopolymeric tract organization in *P. falciparum* and other Apicomplexan parasites.

Early promoter deletion studies suggested that the removal of poly dA.dT tracts in upstream flanking sequence alters absolute levels of gene expression (Porter, 2002; Polson and Blackman, 2005). The evaluation of simple sequence repeats in 27 eukaryotic organisms, including *P. falciparum* (chromosome II and III), using a quantitative algorithm called Poly (Bizzaro and Marx, 2003) identified that long poly dA.dT tracts, for most organisms investigated, were over-represented within non-coding sequence and significantly enriched at >4-10bp lengths (Zhou *et al.*, 2004). Moreover, the length of these tracts was over-proportional, i.e. longer than expected by chance, particularly in organisms of 30-50% genomic GC content (Zhou *et al.*, 2004). For *P. falciparum*, poly dA.dT tracts were also observed to be over-represented within coding sequence (Zhou *et al.*, 2004).

A subsequent study (Cohanin and Haran, 2009) investigated the relationship between nucleosome positioning and incidence of genomic homopolymeric poly dA tracts in *S.*

cerevisiae, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Homo sapiens* and *Danio rerio*. This study identified that the use of local codon bias was likely responsible for long homopolymeric poly d.A tract avoidance and that poly d.A tracts were more prolific within intergenic regions - particularly immediately 5' and 3' of the ORF and introns. Correlating the poly d.A incidence of the less densely populated exonic regions with a *S. cerevisiae* microarray-based nucleosome map this group identified that there was also a distinct difference in nucleosome avoidance depending upon the length of the homopolymeric A tract - whereupon only tracts of length $A_{n>5}$ appeared nucleosome refractory (Cohanin and Haran, 2009). Short homopolymeric dinucleotide repeats AA/TT/TA, with a 10 base pair periodicity, are implicated in DNA bending and hence nucleosome formation (Anselmi *et al.*, 1999). Longer poly dA.dT tracts are thought to be nucleosome refractory owing to their intrinsic DNA structure (Woods *et al.*, 2004). The hydrated narrow minor groove and shorter helical repeat of poly dA.dT DNA appears to cause maximal nucleosome depletion over the tract itself, however, nucleosome depletion can also extend out from the tract for up to ± 100 -150bp and this appears to be related to the perfection and length of the poly dA.dT tract itself (Segal and Widom, 2009b). Interestingly, in accordance with previous speculation, Cohanin and Haran, on the basis of their data, did not support a hypothesis for tract expansion via a strand-slippage mechanism (Zhou *et al.*, 2004; Cohanin and Haran, 2009). Long poly d.A tracts are also known to cause the termination of DNA synthesis in retroviruses by reverse transcriptase (Lavigne and Buc, 1999) and this is related to the structural narrowing of the minor groove in polyA DNA (minimal at 7bp), however, $A_{n>5}$ tracts were not avoided in exonic nucleosome linker regions - only nucleosome-bound exonic regions (Cohanin and Haran, 2009). Therefore, this group hypothesized that poly d.A tracts were more likely nucleosome positioning signals (Cohanin and Haran, 2009).

Studies in yeast identify two types of promoter region, TATA-containing and TATA-less, and correlate nucleosome proximity, transcriptional plasticity and DNA rigidity with these two

types of promoter, respectively (Tirosh *et al.*, 2007; Tirosh and Barkai, 2008). Data from these studies suggest that TATA-less promoters have a broader, species-specific, nucleosome-free region (NFR) and that this region is associated with rigid DNA (Tirosh *et al.*, 2007; Tirosh and Barkai, 2008). Whilst homopolymeric dA.dT tracts were infrequent in TATA-containing promoters, they are frequent, highly correlated with the centre of the NFR and exhibit strand-specific symmetry in TATA-less promoters (Wu and Li, 2010). These data suggest that, in TATA-less promoters, the presence of poly dA.dT tracts may enable the creation of a nucleosome refractory 'access' region around the TSS. Studies in *Dictyostelium discoideum* (another extremely AT-rich organism) have also identified an enrichment of poly dT tracts on the sense strand just upstream of the TSS and an enrichment of poly dA tracts on the sense strand just downstream of the transcription end site (TES) - the positions of which correlate with the presence of NFRs (Chang *et al.*, 2012). In addition, a general enrichment of either poly dA, dT or poly dAT was observed in nucleosomal linker regions and at nucleosomal borders suggesting that homo- or hetero-polymeric tracts may be important in the largely life-cycle invariant chromatin organization of this organism (Chang *et al.*, 2012). A barrier model has also been proposed by Mavrigh *et al.*, who suggested that nucleosomal packing in *S. cerevisiae* could be guided by the robust positioning of the +1 and -1 nucleosomes adjacent to the NFR (Mavrigh *et al.*, 2008).

Nucleosome occupancy studies in *P. falciparum* demonstrate chromosomes with densely packed heterochromatic sub-telomeric and centromeric regions and more relaxed euchromatic chromosomal internal regions, containing nucleosome bound exons and intergenic regions which are populated with an H2A.Z and apicomplexan-specific H2B.Z double nucleosome variant (Bartfai *et al.*, 2010; Hoeijmakers *et al.*, 2013; Bartfai *et al.*, 2013). Nucleosome occupancy appears to vary during the parasite development cycle, with minimal occupancy corresponding to highest levels of transcriptional activity and maximal occupancy levels towards the end of the IE cycle (Westenberger *et al.*, 2009; Ponts *et al.*, 2010). It is

interesting therefore, within this context, that these intergenic histone double-variants (H2Z.A and H2Z.B), which are thought to be less stable than their canonical counterparts, appear to express a preference for association with AT-rich sequences (Hoeijmakers *et al.*, 2013).

In summary, it has been proposed that poly dA.dT tracts may function in diverse roles including: nucleosomal positioning determinants, promoter sequences - which may be synergistic with their role in nucleosome positioning, and/or as binding sites for *trans*-acting factors (Zhou *et al.*, 2004; Chang *et al.*, 2012). Current data suggests that poly dA.dT tracts are indeed not likely themselves to be determinants of transcription, but instead may facilitate transcription by creating a nucleosome free zone which would enable access by *trans*-acting factors to *cis*-acting DNA sequences around the transcription start site (Arya *et al.*, 2010; Wu and Li, 2010).

This chapter uses comparative analysis with the Poly algorithm on a genome wide scale to explore the over-representation and length over-proportionment of poly dA.dT tracts within what may be considered functionally equivalent 5' and 3' flanking sequences in five *Plasmodium spp.* and a range of other Apicomplexan organisms (*Theileria spp.*, *Cryptosporidium spp.*, *Babesia bovis*, *Toxoplasma gondii* and *Neospora caninum*) of varying genomic AT-content. Positional bias of these poly dA.dT tracts, relative to the ORF and putative TSSs (*P. falciparum* only), is explored and correlated with current data on nucleosome positioning.

6.2 COMPARATIVE ANALYSIS OF HOMOPOLYMER TRACT FREQUENCY AND LENGTH IN THE PROXIMAL INTERGENIC REGIONS OF APICOMPLEXAN PARASITES

Poly (www.bioinformatics.org/poly) is a bioinformatic algorithm (Bizzaro and Marx, 2003) that quantitatively assesses the incidence of single sequence repeats (SSR) in any given sequence. It was used to assess the incidence and length of homopolymeric sequences within the 5' and 3' flanking sequence of five *Plasmodium spp.* and eight other Apicomplexan parasites (see Chapter 2 Figure 5 for Poly workflow). Annotated sequence data was downloaded from PlasmoDB.org (*Plasmodium spp.*), Gene DB.org (*Theileria spp.* and *B. bovis*), CryptoDB.org (*Cryptosporidium spp.*) and ToxoDB.org (*T. gondii* and *N. caninum*). The phylogenetic relationship between all organisms investigated here is shown in Fig. 6-1 and the list of all Apicomplexan organisms evaluated is shown in Table 6-1. Using `all.intergenic.dist.2.FASTA.pl`, the chromosomal sequence files and the parsed general feature format (.gff) file, upstream and downstream sequence files were compiled for each organism evaluated. The size of the flanking upstream regions and downstream regions utilized for each organism were taken from Chapter 3 results (c.median size of the A and C type IGR respectively - data shown in Table 6-1). It should be noted that not all individual sequences within each dataset were of the full window length. This was attributable to upstream and downstream ORF encounters by `intergenic.dist.2.FASTA.pl` whereby the intergenic sequence becomes truncated at this point (as this sequence is no longer intergenic). The sequence data were converted into a Poly compatible format using `fasta2poly.pl` - an individual sequence tagging process to prevent 'artefacts' arising from two joined sequences - and each sequence was run through `Poly.py` (Bizzaro and Marx, 2003).

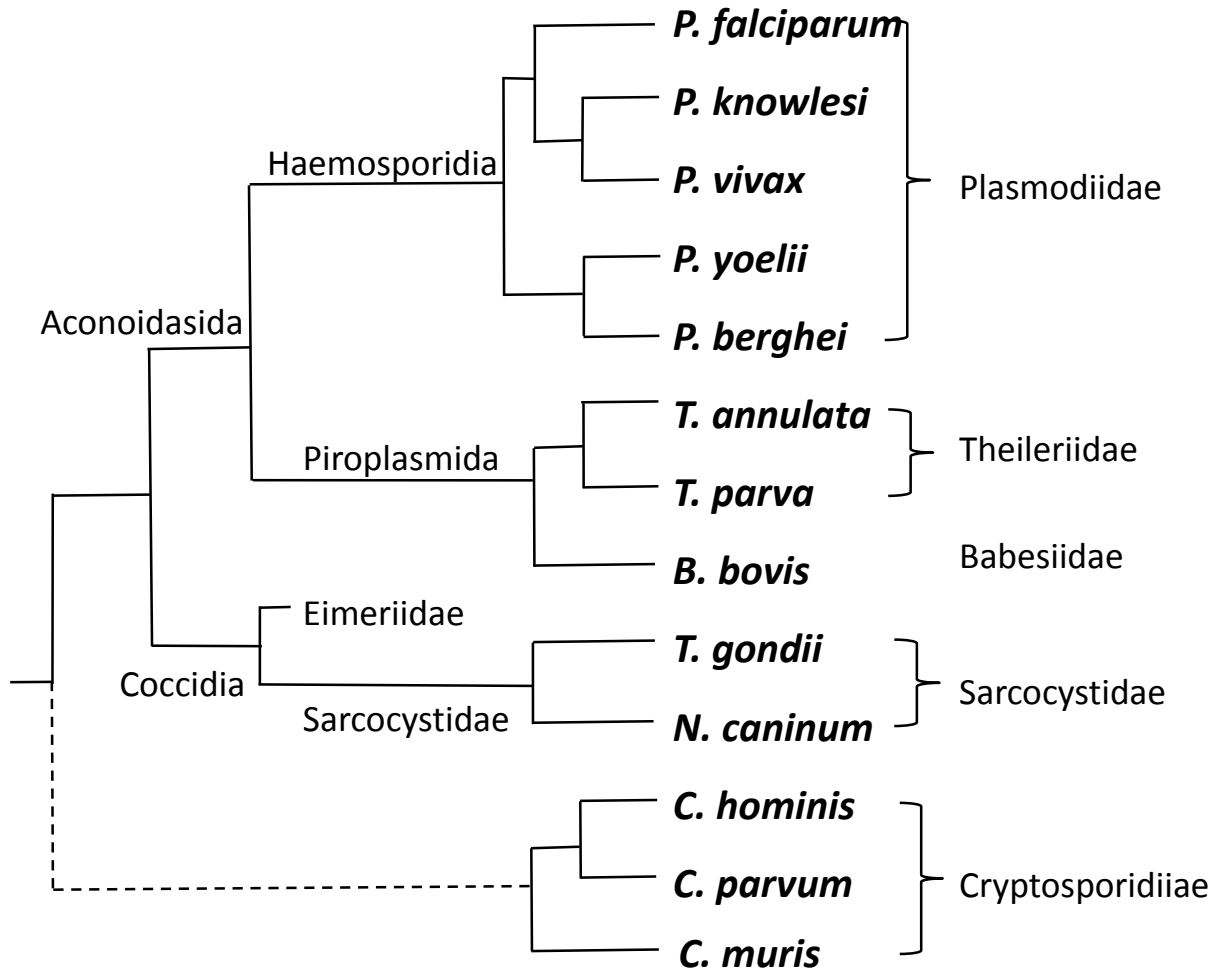


Figure 6-1 Phylogenetic relationship between Apicomplexan parasites used in this study

This figure demonstrates the phylogenetic relationship between the 13 organisms investigated during this study. As such it should be noted that the branches of the tree are incomplete. *Cryptosporidium spp.* lack the apicoplast organelle and hence the undefined relationship of these early branching Apicomplexan parasites is indicated with a dotted line. Organism orders and sub-orders are indicated to the left and families to the right.

Table 6-1 Genomic AT content, gene density and IGR window size for each Apicomplexan organism evaluated

Organism	Family	% AT ¹	Gene Density ²	Median size of IGR ³ (bp)			Ratio of IGR size ⁴			Ups (bp) ⁵	Down (bp) ⁵
				A	B	C	A	B	C		
<i>Plasmodium falciparum</i>	Plasmodiidae	80.6	4.3	1938	1385	677	2.9	2.0	1.0	2000	700
<i>Plasmodium knowlesi</i>	Plasmodiidae	62.5	4.6	2162	1592	736	2.9	2.2	1.0	2000	700
<i>Plasmodium vivax</i>	Plasmodiidae	57.7	4.5	1956	1434	643	3.0	2.2	1.0	2000	700
<i>Plasmodium yoelli</i>	Plasmodiidae	77.4	2.6	1192	578	582	2.0	1.0	1.0	1000	700
<i>Plasmodium berghei</i>	Plasmodiidae	79.2	3.1	na	na	na	na	na	na	1000	700
<i>Cryptosporidium hominis</i>	Cryptosporidiidae	68.3	2.3	640	494	203	3.2	2.4	1.0	650	200
<i>Cryptosporidium parvum</i>	Cryptosporidiidae	70.0	2.4	634	460	175	3.6	2.6	1.0	650	200
<i>Cryptosporidium muris</i>	Cryptosporidiidae	83.0	na	na	na	na	na	na	na	650	200
<i>Toxoplasma gondii</i>	Sarcocystidae	47.7	9.1	2576	2437	1623	1.6	1.5	1.0	2500	1600
<i>Neospora caninum</i>	Sarcocystidae	45.2	8.6	3603	3899	2172	1.7	1.8	1.0	2500	1600
<i>Babesia bovis</i>	Babesiidae	58.2	2.2	543	352	175	3.1	2.0	1.0	550	200
<i>Theileria annulata</i>	Theileriidae	67.5	2.2	439	277	125	3.5	2.2	1.0	450	150
<i>Theileria parva</i>	Theileriidae	65.9	2.1	376	256	154	2.4	1.7	1.0	450	150

¹%AT content of whole genome. ²As total genome/ number of genes in kbp. ³InterGenic Region. ⁴Where size of IGR C is defined at 1. ⁵Length of proximal upstream (Ups) and downstream (Down) sequence investigated here.

Poly essentially uses a window size of $n=1$ to assess each subsequent adjacent base in a given sequence to establish whether it is the same or different from the previous base. This way it records, for each base type; (i) the homopolymeric tract incidence (c) at length N (c_{iNobs}) for each nucleotide within the input sequence (l_{seq}). Therefore, the observed tract frequency (f_{iNobs}) is denoted as:

$$f_{iNobs} = c_{iNobs} / l_{seq}$$

From the input sequence an expected frequency of the random occurrence for base i of length N for this particular input sequence is calculated (f_{iNexp}) (Bizzaro and Marx, 2003) and is denoted as:

$$f_{iNexp} = f_{i1obs}^N \times (1 - f_{i1obs})^2$$

The actual Representation (R) of that particular tract length for that particular base is a ratio of expected (f_{iNexp}) v observed (f_{iNobs}) and due to the typical extent of overrepresentation is described using the \log_{10} function. It is denoted as:

$$R = \log_{10}(f_{iNobs} / f_{iNexp})$$

Each base output file contains the degree of representation (R) for that particular base of that particular length for that particular input sequence where R values larger than 0 are considered over-represented and R values of less than 0 are under-represented. Using $R > 0.5$ (*i.e.* at least 3.16-fold over-represented) to provide a defined threshold for tract length over-representation for each base (Bizzaro and Marx, 2003). A linear plot of N at $\log(R)$ denotes slope_R to describe the extent of the over-representation. This slope is determined between tract lengths where $R > 0.5$ and N_{maxobs} (using $c_{iNobs} \geq 4$ to reduce noise associated with infrequent observations) - an example of this type of plot is shown in Fig. 6-2.

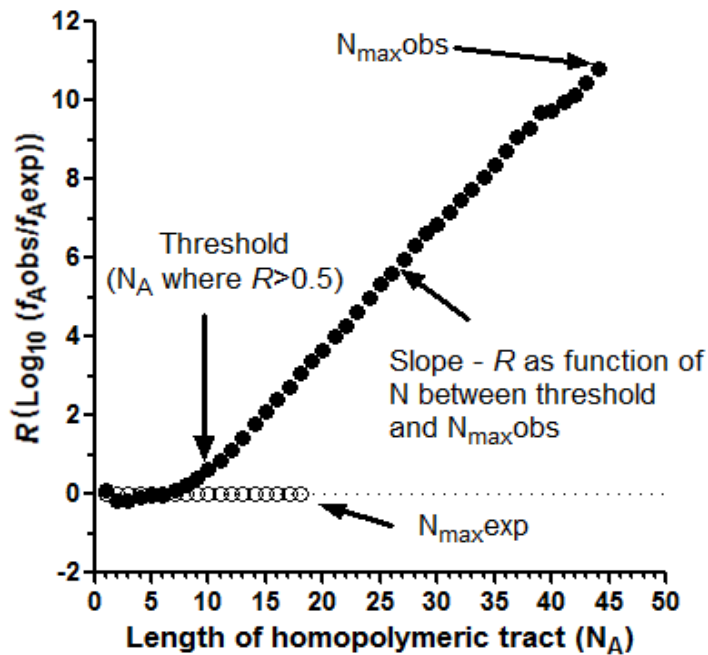


Figure 6-2 Example output from Poly analysis of homopolymer tract frequency

This figure represents an example of poly output for *P. falciparum* upstream sequence (2000bps) base A where \log_{10} representation (R) (f_{Aobs}/f_{Aexp}) is plotted as a function of length (N_A). N_{maxexp} represents the maximal length of base A predicted from the randomised input sequence (Bizzaro and Marx, 2003) and is represented by unfilled circles. N_{maxobs} is the maximal length of base A actually observed in the input sequence where $N > 4$ and is represented by filled circles. The threshold of over-representation occurs when N_A of f_{Aobs}/f_{Aexp} reaches $R > 0.5$. The $slope_R$ is determined from the point of the departure threshold and N_{maxobs} .

Proportionment (P) is an evaluation of tract length where P values larger than 1 are considered over-proportional and values of less than 1 under-proportional. P is calculated using N_{maxexp} and N_{maxobs} values whereby N_{maxexp} is calculated from the input sequence (Bizzaro and Marx, 2003) and N_{maxobs} is the actual maximum tract length observed for a particular base within that particular input sequence (where $c_{iNobs} \geq 4$). Therefore, P is denoted by;

$$P = N_{maxobs}/N_{maxexp}$$

In addition to the four output A, T, G and C base files created for each input sequence file a further supplementary output file is produced which records the total number of bases per

sequence analysed, the frequency of each base (as a % of total bases in this particular sequence), the % AT/GC composition, the $N_{\max\text{exp}}$ and threshold length N for $R>0.5$.

Therefore, Poly was used to provide, for each input sequence, the fraction (Fraction_i) of each nucleotide, the incidence of each nucleotide at length N ($f_{N\text{obs}}$) and the maximal length of each nucleotide tract ($N_{\max\text{obs}}$). For each input sequence and each nucleotide these data were then compared with their cognate $f_{N\text{exp}}$ and an $N_{\max\text{exp}}$ - created by Poly from the randomized input sequence (Bizzaro and Marx, 2003). This enabled the degree of representation (R) for each tract and the proportionment (P) (relative length) of the actual tracts for each base in each sequence to be ascertained. Linear interpolation of the expected v observed data in the form of $\text{Log}_{10}(f_{A/T/G/C\text{obs}}/f_{A/T/G/C\text{exp}})$ against N_i (see Fig. 6-2) using $R>0.5$ as the threshold (therefore the observed frequency is 10^5 - 3.16-fold higher than expected by chance) allows the departure point from, and degree of, over-representation to be evaluated. The N_i endpoint of each curve, the difference between the expected length of the homopolymeric tract ($N_{\max\text{exp}}$) and the actual observed $N_{\max\text{obs}}$ (where $N_{\max\text{obs}} \geq 4$), enables the degree of over-proportionment to be identified.

A linear interpolation of the Poly output for the ORF upstream and downstream regions for *Plasmodium spp.* and *Cryptosporidium spp.* is shown in Fig. 6-3. What is apparent for these representatives of the Plasmodiidae and Cryptosporidiidae families is a clear over-representation of homopolymeric tracts for all base types (A/T/G and C) within both ORF-adjacent intergenic compartments; in particular, short polydG.dC tracts and long homopolymeric dA.dT tracts.

Figure 6-3 Representation of homopolymeric tracts in the proximal upstream and downstream intergenic flanking regions of *Plasmodium spp.* and *Cryptosporidium spp.*

(overpage) Graphs plotting R against N_i for Bases A (red), T (blue), G (black) and C (green) for the group of organisms that demonstrated overrepresentation of short poly dG.dC tracts and long poly dA.dT tracts in their proximal upstream and downstream intergenic flanking sequence. The %AT content for each dataset is shown on each graph.

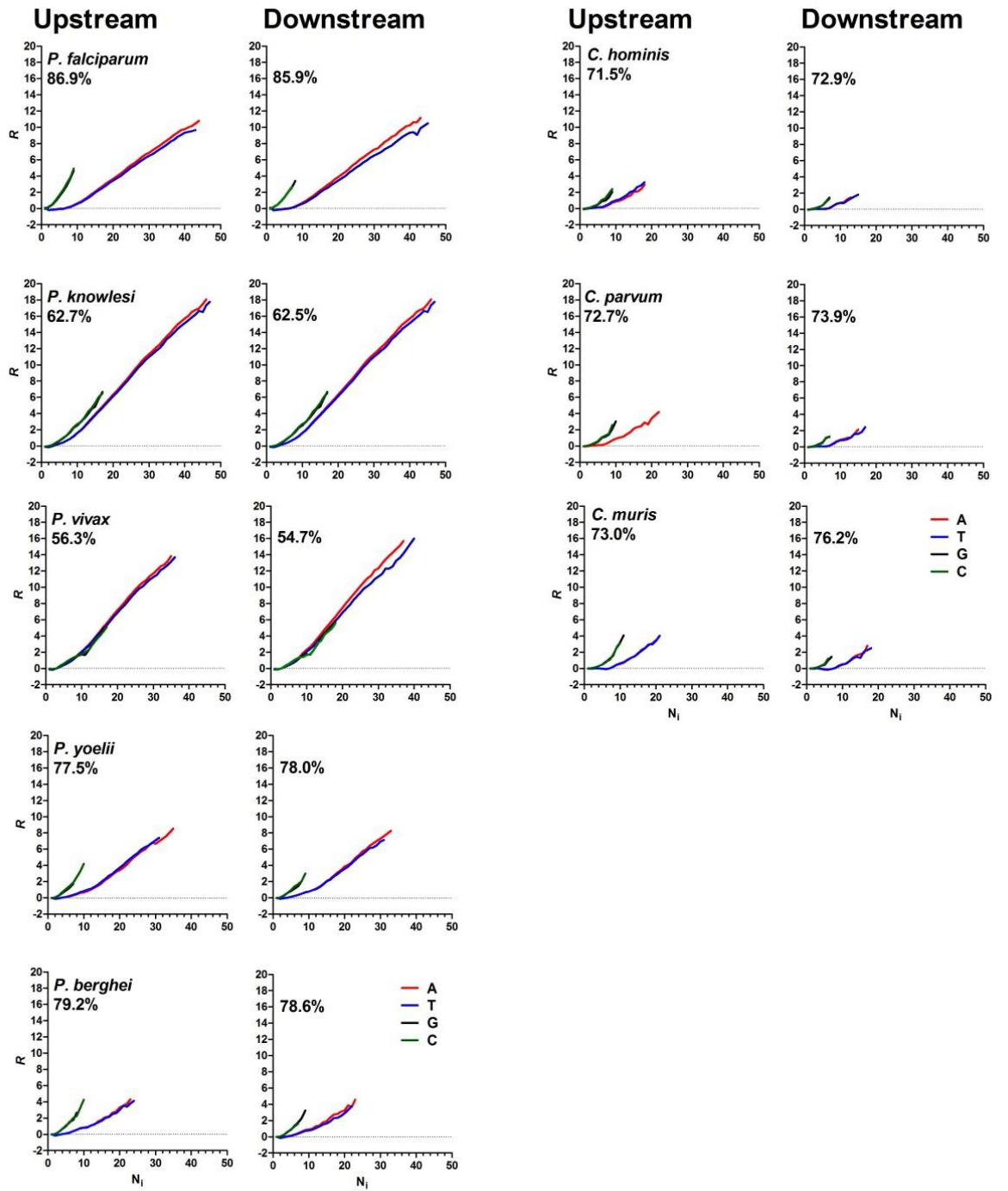


Figure 6-4 shows a linear interpolation of the Poly output for the ORF upstream and downstream regions for representatives of the Theileriidae, Babesiidae and Sarcocystidae families. For *T. gondii* and *N. caninum* (Sarcocystidae) an over-representation of homopolymeric dA.dT.dG and dC tracts is again apparent but there is a reversal in their relative length with poly dA.dT tracts shorter and poly dG.dC tracts longer in these organisms, though not as long as those observed for *Plasmodium spp.* Interestingly, neither of these two compartments in *Theileria spp.* or *B.bovis* demonstrated any homopolymeric tract over-representation. Therefore, the Apicomplexan organisms investigated here seem to represent three distinct groups, each with different homopolymeric tract content residing within their 5' and 3' ORF adjacent flanking regions. Critically, closely related organisms share the same organisation of homopolymeric tracts within these intergenic compartments.

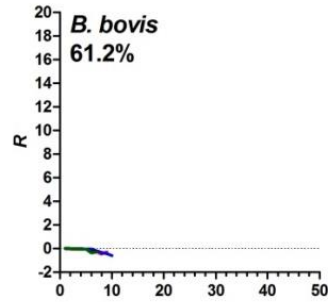
Poly analysis of the coding sequence (CDS) for this group of organisms again produces a very pronounced grouping (See Fig. 6-5). Whereas, homopolymeric dA.dT.dG and dC tracts overrepresentation is virtually absent from all other organisms investigated (*Homo sapiens* and *Mus musculus* are also included for comparative purposes) they are abundant within the coding sequence of all *Plasmodium spp.* with a similar pattern to that identified within ORF adjacent upstream and downstream regions; that is short homopolymeric poly dG.dC and long poly dA.dT tracts. It is interesting to note the length of the poly dA.dT tracts within CDS for the human infective *Plasmodium spp* (*P. falciparum*, *P. vivax* and *P. knowlesi*) - all with varying underlying base content and the trend towards a higher over-representation of Poly dT. However, it should also be noted that the base sequence composition of CDS is inherently skewed (*P. falciparum* 44.58% A, 30.92% T, 14.2% G and 10.08% C) whereas this is not the case in upstream or downstream compartments where A to T and G to C base content are proportional. Therefore, within the context of a lower ORF base T content, poly dT tracts would appear to be proportionally over-represented. Longer lengths of poly dA tracts are also found in the two mouse malaria parasites (*P. yoelii* and *P. berghei*) but are not as

prevalent as in the above *Plasmodium spp.*, – perhaps not unexpected due to the shorter lengths of IGR analysed in these two organisms.

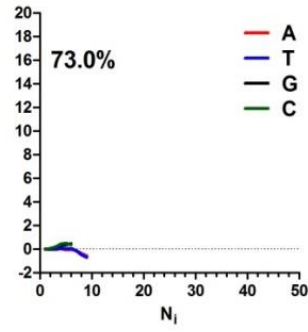
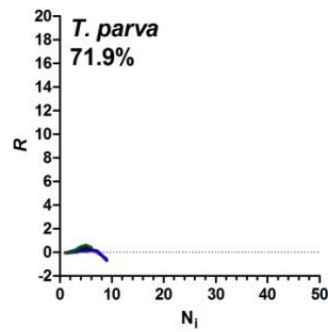
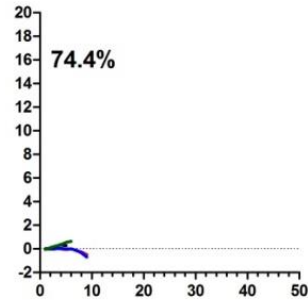
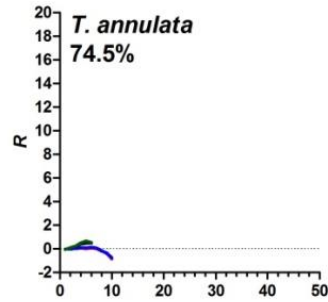
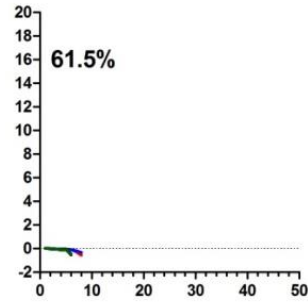
Figure 6-4 Representation of homopolymeric tracts in the proximal upstream and downstream intergenic flanking regions of the coccidian and piroplasmida organisms

(overpage) Graphs plotting R against Ni for Bases A (red), T (blue), G (black) and C (green) for the groups of organisms that demonstrated either; (A) no evidence of homopolymeric tract overrepresentation (*Theileria spp.* and *B. bovis*) or (B) long poly dG.dC tracts and short poly dA.dT tracts overrepresentation (*T. gondii* and *N. caninum*). The %AT content for each dataset is shown on each graph.

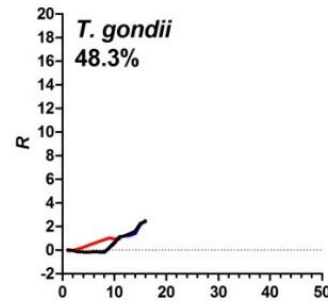
A Upstream



Downstream



B Upstream



Downstream

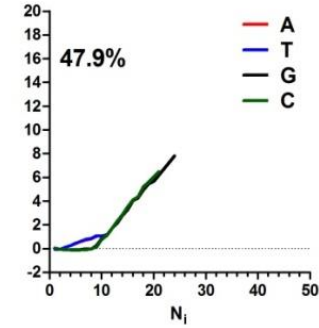
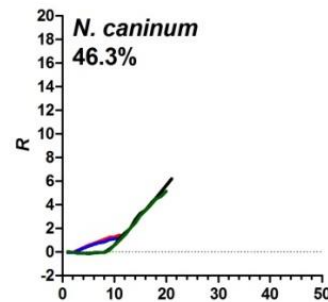
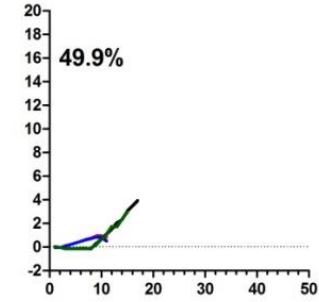


Figure 6-5 Representation of homopolymeric tracts in coding sequence of the coccidian and piroplasmida organisms.

(overpage) Graphs plotting R against Ni for Bases A (red), T (blue), G (black) and C (green) for the CDS of twelve of the Apicomplexan organisms previously investigated. Unfortunately *N. caninum* CDS was unavailable at the time of analysis and is therefore not included. Poly analysis of the CDS for *H. sapiens* and *M. musculus* is included for comparative purposes. The %AT content for each dataset is shown on each graph.

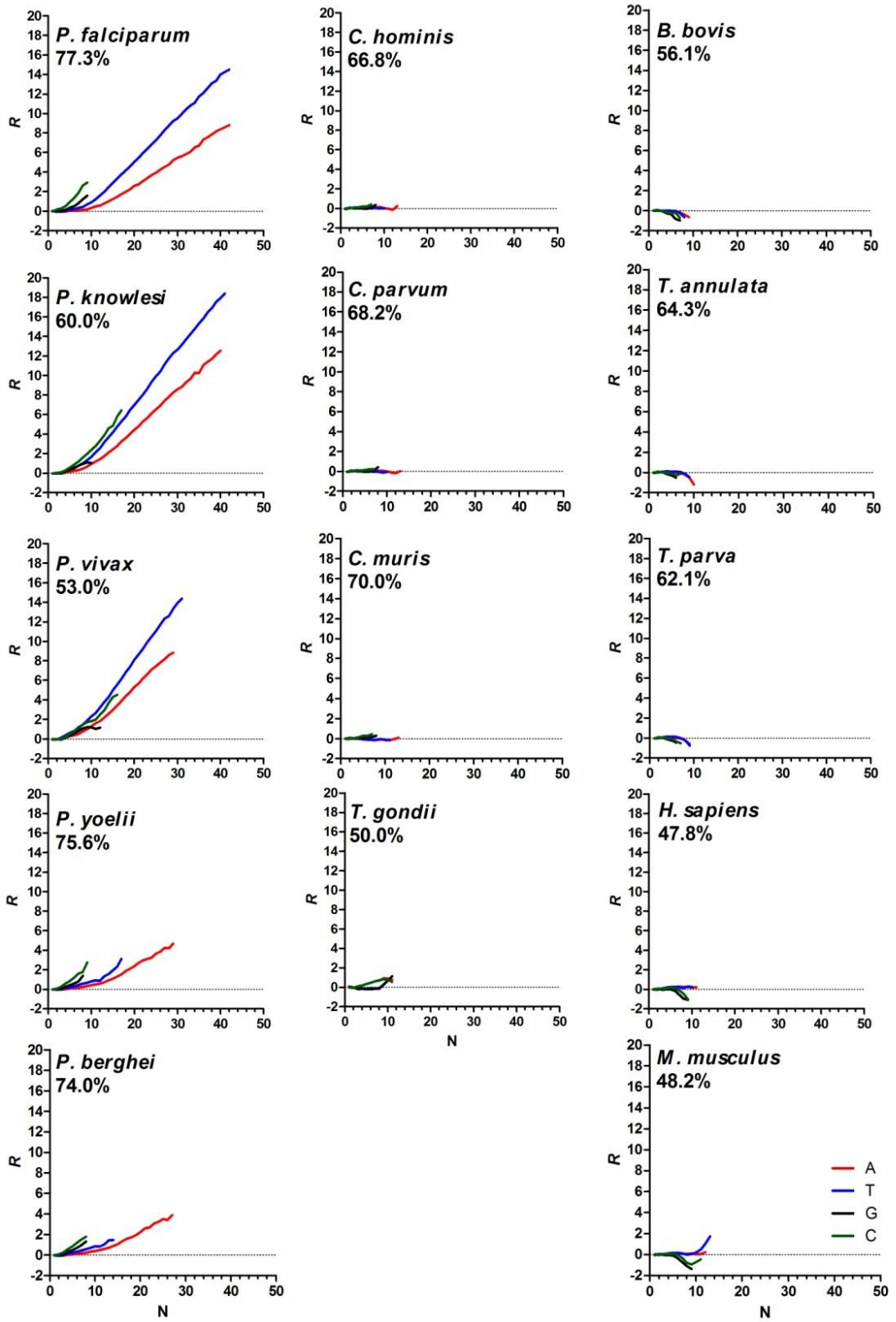


Table 6-2 shows data for each base from each organism for the 5' and 3' flanking regions namely; the proportionment (P) (derived from $N_{\max\text{Obs}}/N_{\max\text{Exp}}$), the threshold of over-representation (R) (or departure point derived from $f_{\text{NObs}}/f_{\text{NExp}}$ where $R>0.5$) and the slope_R (the slope of over-representation derived via linear regression of the $N_{\max\text{Exp}}$ to $N_{\max\text{Obs}}$ curve). Using data from Table 6-2 various parameters were investigated to correlate; i) nucleotide composition with homopolymeric tract occurrence and length and ii) explore the relationship between nucleotide composition, homopolymeric tract length and the size of the flanking intergenic region.

Comparative analyses of the relative level of over-representation between species for homopolymeric tracts of different base composition (A, T, G or C) were determined using linear regression of slope_R against Fraction_i and threshold against Fraction_i . Whilst it would be anticipated that slope_R would decrease with a higher individual base content and hence a higher f_{NExp} value this does not appear to be the case. Although a weak general linear correlation was observed for all bases ($r^2 = 0.33$, $p<0.001$, slope = 0.8 data not shown) the slope_R for individual bases (apart from base C in the upstream region) do not appear to have a linear correlation with the base composition (Fig. 6-6A). This suggests that the degree of homopolymeric tract over-representation within these compartments is not related to the underlying base composition. However, a good correlation is observed for all bases from the different species when the threshold, rather than slope_R , are considered (Fig. 6-6B). These data are in accordance with Zhou *et al.* who identified that as the base composition increased so too did the threshold value of N_i (Zhou *et al.*, 2004). Therefore, the length of the homopolymeric tract (N_i), before it reaches its departure point from expected for the input sequence, does appear to be related to the base composition. That is, the dynamic threshold for tract expansion is not a simple reflection of length alone – but instead depends upon the underlying base composition.

Table 6-2 Upstream and downstream values for Fraction_i, N_{maxObs}, N_{maxExp}, P, R and slope_R for all Apicomplexan organisms investigated

Organism	Fraction _i				base T	Maximum length observed N _{maxObs}				Maximum length expected M _{maxExp}				Proportionment (P)				Threshold (R>0.5)				Slope _R ¹			
	frequency					A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
	A	C	G	T																					
Upstream																									
<i>P. falciparum</i>	0.43	0.06	0.07	0.44	44	9	9	43	18	5	5	19	2.44	1.80	1.80	2.26	10	3	4	10	0.31	0.72	0.74	0.29	
<i>P. knowlesi</i>	0.30	0.19	0.18	0.31	46	17	17	47	13	9	9	13	3.54	1.89	1.89	3.62	7	5	5	7	0.46	0.49	0.48	0.44	
<i>P. vivax</i>	0.27	0.22	0.21	0.29	35	17	16	36	12	10	10	12	2.92	1.70	1.60	3.00	6	5	5	6	0.47	0.36	0.35	0.45	
<i>P. yoelii</i>	0.42	0.10	0.12	0.36	35	10	7	32	18	7	7	16	1.94	1.43	1.00	2.00	10	4	5	8	0.32	0.57	0.40	0.32	
<i>P. berghei</i>	0.39	0.10	0.10	0.40	23	10	7	24	16	6	6	16	1.44	1.67	1.17	1.50	8	4	4	8	0.25	0.57	0.48	0.23	
<i>C. hominis</i>	0.36	0.14	0.15	0.35	18	9	9	18	14	7	7	13	1.29	1.29	1.29	1.38	9	5	6	8	0.24	0.46	0.39	0.26	
<i>C. parvum</i>	0.37	0.13	0.14	0.36	22	9	10	na ²	14	7	7	na	1.57	1.29	1.43	na	9	5	5	na	0.26	0.45	0.47	na	
<i>C. muris</i>	0.36	0.13	0.14	0.37	20	10	11	21	14	7	7	14	1.43	1.43	1.57	1.50	10	6	6	10	0.30	0.60	0.67	0.30	
<i>T. gondii</i>	0.23	0.26	0.26	0.25	13	17	17	na	11	12	12	na	1.18	1.42	1.42	na	6	10	10	na	0.08	0.29	0.29	na	
<i>N. caninum</i>	0.22	0.27	0.26	0.25	11	20	21	11	10	12	12	11	1.10	1.67	1.75	1.00	5	10	10	6	0.14	0.47	0.51	0.10	
<i>B. bovis</i>	0.31	0.19	0.20	0.30	9	7	7	10	12	8	8	12	0.75	0.88	0.88	0.83	na	na	na	na	na	na	na	na	
<i>T. annulata</i>	0.37	0.13	0.13	0.37	11	6	6	11	14	6	7	14	0.79	1.00	0.86	0.79	na	4	7	na	na	na	na	na	
<i>T. parva</i>	0.36	0.14	0.15	0.35	9	6	6	9	14	7	7	13	0.64	0.86	0.86	0.69	na	5	na	na	na	na	na	na	
Downstream																									
<i>P. falciparum</i>	0.41	0.07	0.07	0.44	43	7	8	45	17	5	5	18	2.53	1.40	1.60	2.50	10	4	4	10	0.32	0.59	0.63	0.29	
<i>P. knowlesi</i>	0.29	0.19	0.19	0.33	42	16	16	44	12	9	8	13	3.50	1.78	2.00	3.38	6	5	5	7	0.46	0.49	0.48	0.44	
<i>P. vivax</i>	0.26	0.23	0.22	0.29	32	15	16	32	11	10	10	11	2.91	1.50	1.60	2.91	6	5	5	6	0.50	0.39	0.40	0.45	
<i>P. yoelii</i>	0.38	0.12	0.11	0.39	33	9	8	31	15	6	6	16	2.20	1.50	1.33	1.94	8	4	4	8	0.33	0.45	0.34	0.31	
<i>P. berghei</i>	0.38	0.11	0.10	0.41	24	7	9	22	14	6	6	16	1.71	1.17	1.50	1.38	8	4	4	8	0.25	0.34	0.50	0.22	
<i>C. hominis</i>	0.35	0.14	0.13	0.38	13	7	7	16	12	6	6	13	1.08	1.17	1.17	1.23	9	6	6	9	0.20	0.42	0.49	0.20	
<i>C. parvum</i>	0.35	0.13	0.13	0.39	15	7	7	19	12	6	6	14	1.25	1.17	1.17	na	9	6	5	9	0.21	0.37	0.36	0.21	
<i>C. muris</i>	0.37	0.12	0.12	0.39	17	6	7	18	13	6	6	14	1.31	1.00	1.17	1.29	10	5	6	11	0.30	0.65	0.50	0.26	
<i>T. gondii</i>	0.25	0.25	0.25	0.25	11	15	17	11	11	11	11	12	1.00	1.36	1.55	na	7	10	10	7	0.06	0.43	0.47	0.01	
<i>N. caninum</i>	0.24	0.25	0.27	0.24	11	21	24	11	11	11	12	11	1.00	1.91	2.00	1.00	6	10	10	6	0.11	0.54	0.51	0.11	
<i>B. bovis</i>	0.30	0.19	0.19	0.31	9	7	7	8	11	8	8	11	0.82	0.88	0.88	0.73	na	na	na	na	na	na	na	na	
<i>T. annulata</i>	0.36	0.12	0.13	0.38	10	7	6	9	13	6	6	13	0.77	1.17	1.00	0.69	na	5	na	na	na	na	na	na	
<i>T. parva</i>	0.35	0.13	0.14	0.38	9	7	6	9	12	6	6	13	0.75	1.17	1.00	0.69	na	5	na	na	na	na	na	na	

¹ slope of R between threshold of overrepresentation and N_{maxObs}. ²na, not available

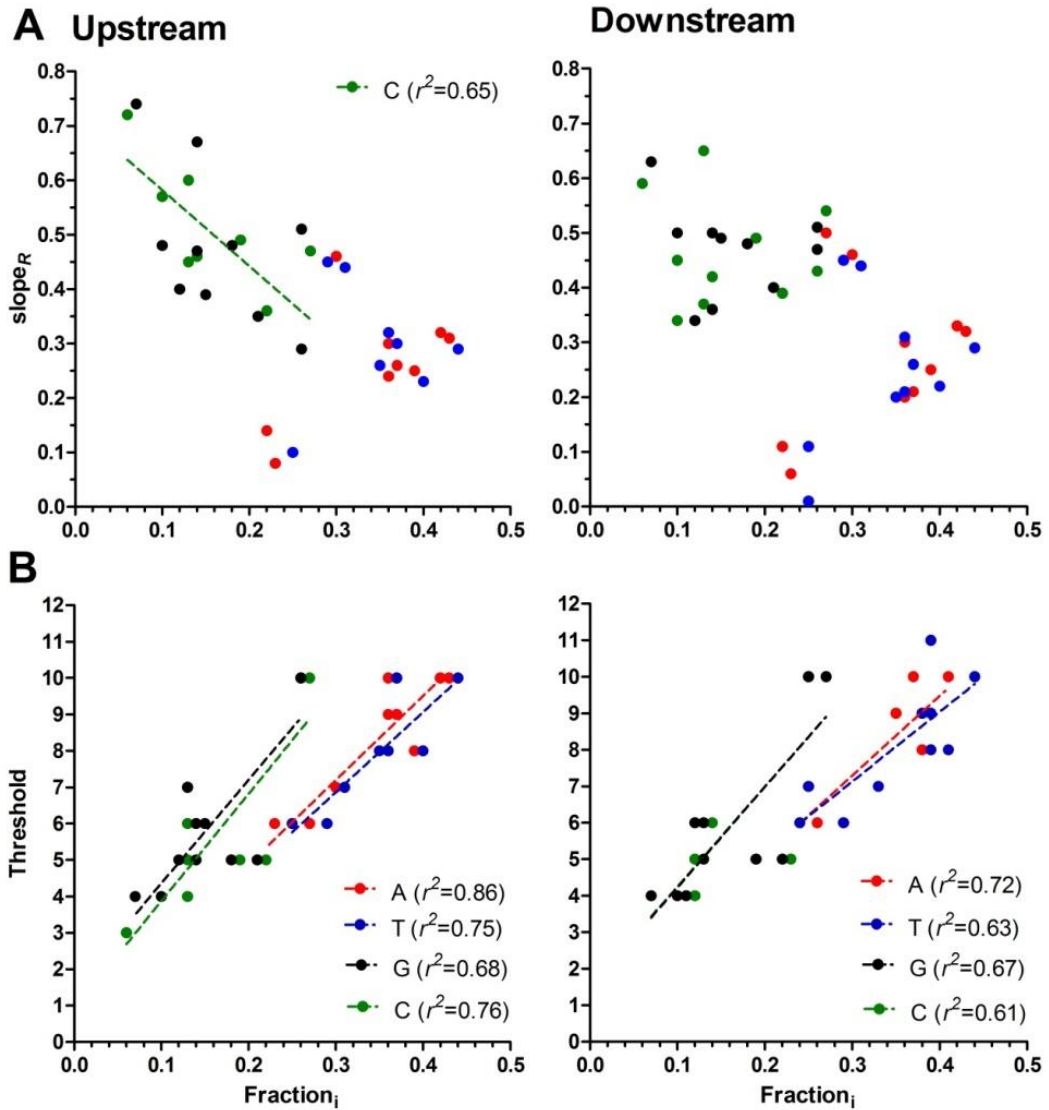


Figure 6-6 Comparative analysis of overrepresentation of homopolymeric tracts as a function of nucleotide content

(A) Slope_R (the slope of overrepresentation) of each nucleotide for each organism is plotted as a function of the fraction of each nucleotide (Fraction_i) for upstream and downstream proximal regions. Only a weak correlation was observed for base C in the upstream proximal region. (B) The threshold of the homopolymeric tract for each nucleotide for each organism is plotted as a function of Fraction_i for upstream and downstream proximal regions. Dotted linear regression lines are shown and correlation coefficients (r^2) are reported where significant. Base A (red), T (blue), G (black) and C (green).

Next, the maximal length of the homopolymeric tract was considered both as a function of base composition and intergenic sequence length. Whereas it would be expected that the maximal length ($N_{\max\text{obs}}$) of a homopolymeric tract would increase with increased base content and sequence length, a moderate correlation only was observed between $N_{\max\text{obs}}$ and base composition for bases G and C. No correlation was observed for bases A and T (Fig. 6-7A). In addition, there was no correlation for any base when P ($N_{\max\text{obs}}/N_{\max\text{exp}}$) was correlated with base content (Fig. 6-7B). Therefore, maximal tract length, $N_{\max\text{obs}}$, for homopolymeric dA or dT tracts, or P for poly dA.dT does not appear to be related to the input sequence base composition for the organisms investigated here. However, both $N_{\max\text{obs}}$ and P , for all base types, correlated well with the input sequence length (the median size of the IGR) (Fig 8A and 8B respectively). It should be noted that poly dA.dT points for *T. gondii* and *N. caninum* were omitted from the regression analysis (marked by *) as previous work (Chapter 3) has demonstrated that these two coccidian parasites with much larger genomes (64 and 59.1Mb respectively) have a different intergenic region size pattern than the 3:2:1 A:B:C ratio exhibited by all the other parasites with compact genomes investigated here (genome densities in the range of 2.0-4.8kb/ORF).

Together these data suggest that the degree of homopolymeric tracts representation in these Apicomplexan genomes is not related to the underlying base composition; however, the threshold for overrepresentation is (defined as $R>0.5$). The maximum length of poly dA.dT tracts also appears unrelated to the underlying base composition, instead, for compact genomes, a good correlation with the length of the IGR is observed. This could perhaps explain why *P. knowlesi* which has a much lower AT- content (61% upstream and 62% downstream) than *P. falciparum* (87% upstream and 85% downstream) but median IGR of equivalent size to *P. falciparum* (*P. knowlesi* A=2162bp and C=736bp, *P. falciparum* A=1938 and C=677) contains equivalent long over-represented homopolymeric dA.dT tracts but with a lower N_i threshold value (*P. falciparum* - R for A/T=10 whereas *P. knowlesi* - R for A /T= 7).

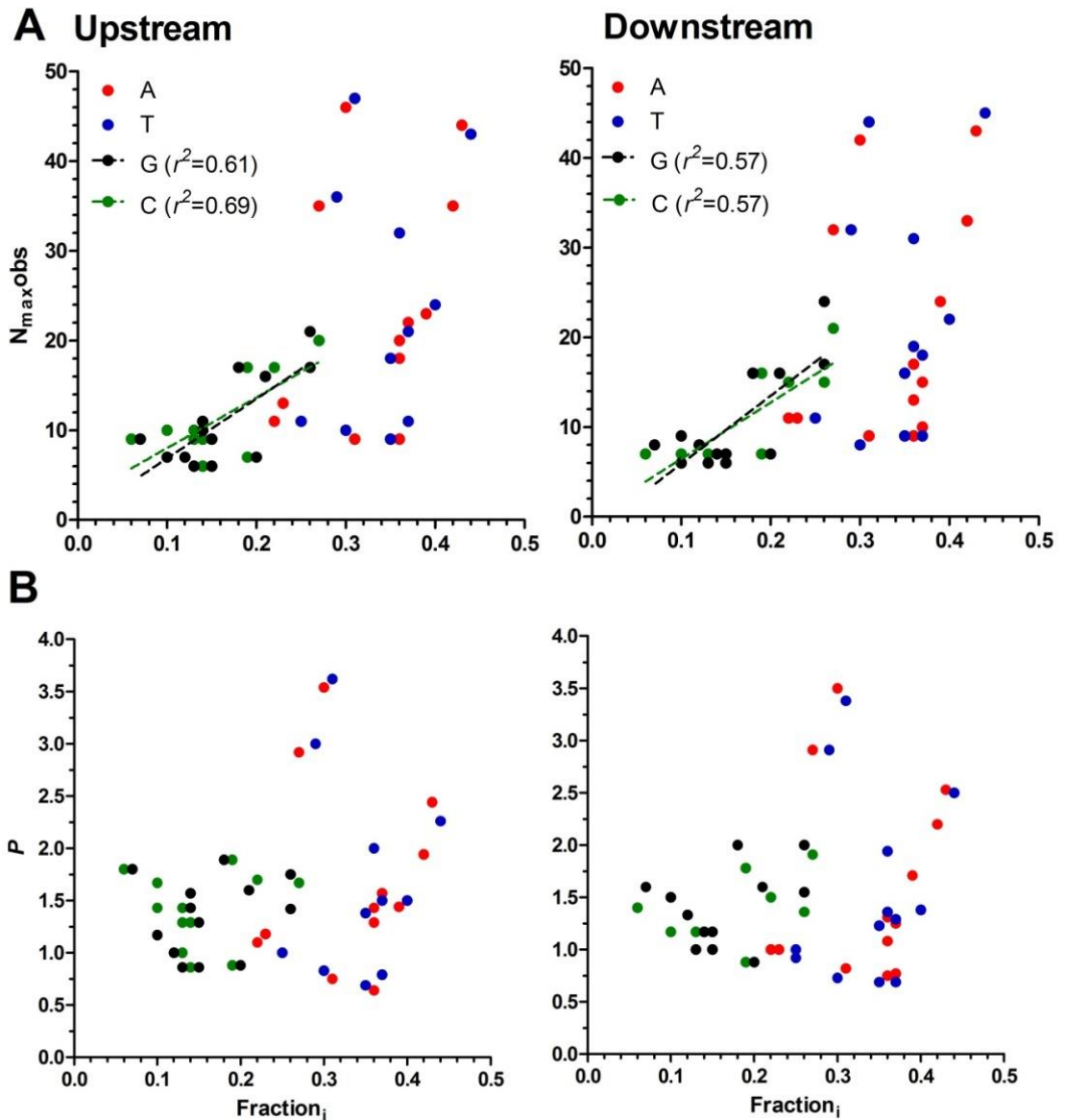


Figure 6-7 Comparative analysis of overproportionment of homopolymeric tracts as a function of nucleotide content

(A) $N_{\max\text{obs}}$ of each nucleotide for each organism is plotted as a function of Fraction_i for upstream and downstream proximal flanking regions. Only poly dG and dC tracts show a significant linear correlation. Dotted linear regression lines and correlation coefficients (r^2) are reported where significant. (B) Proportionment (P) of each nucleotide for each organism as a function of Fraction_i for upstream and downstream proximal flanking regions. No significant linear correlations were observed. Base A (red), T (blue), G (black) and C (green).

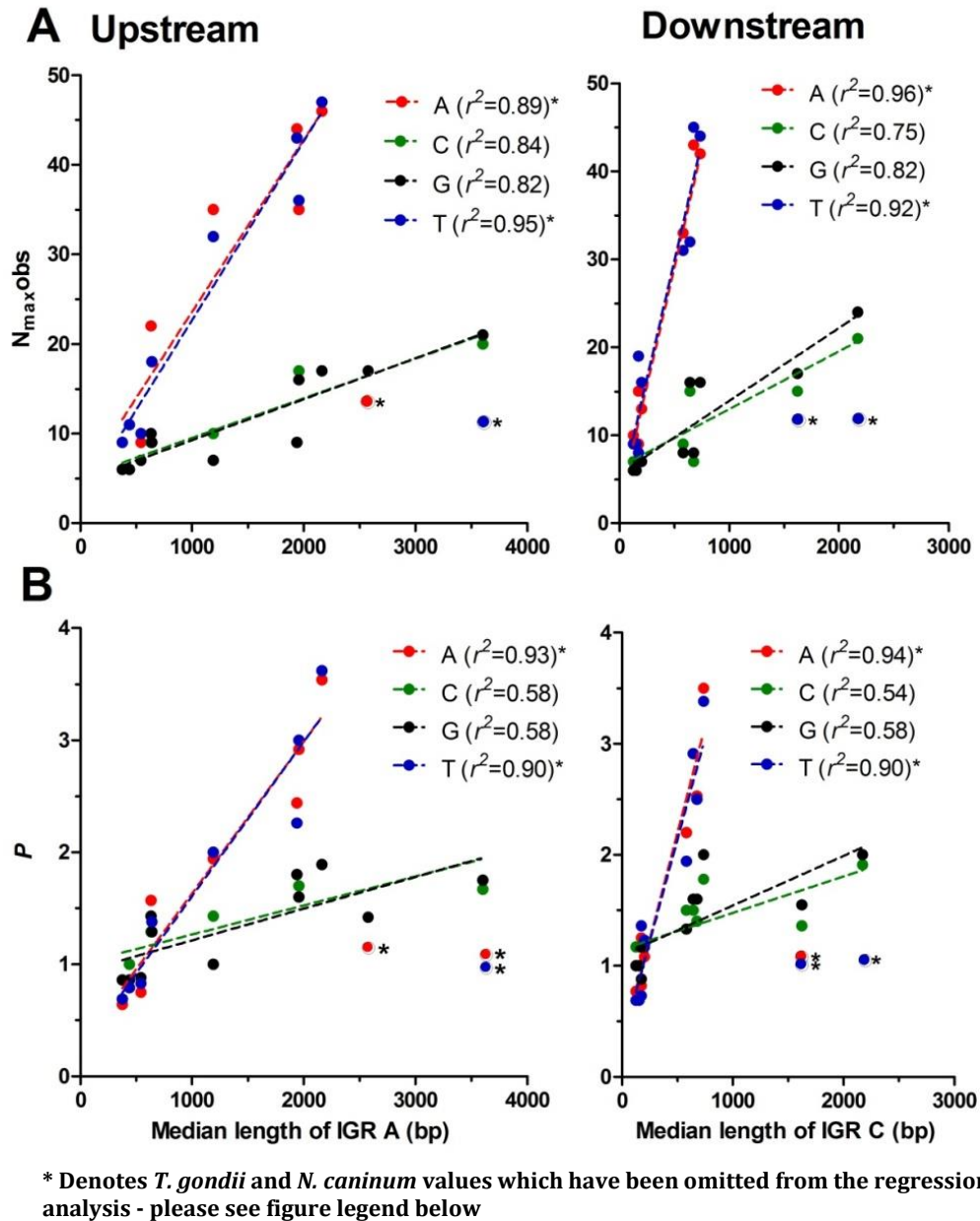


Figure 6-8 Comparative analysis of overproportionment of homopolymeric tracts as a function of size of intergenic region (IGR)

Base A (red), T (blue), G (black) and C (green). Dotted linear regression lines and correlation coefficients (r^2) are reported where significant. (A) $N_{\max\text{Obs}}$ of each nucleotide for each organism as a function of the median length of upstream (type A IGR) and downstream (type C IGR) proximal flanking regions (B) Proportionment (P) of each nucleotide for each organism as a function of the median IGR length (bp) of upstream and downstream proximal flanking regions. Note: *T. gondii* and *N. caninum* values do appear on both the A and B plots but are not included in the regression analysis and are denoted by *. The two coccidian parasites, both of which have much larger genome sizes (64 and 59.1Mb respectively) and much lower gene densities (7.4 and 8.1kb/ORF respectively), were omitted from these analysis as they do not conform to the IGR spacing rule applicable to the other compact Apicomplexan genomes (Chapter 3).

6.3 SPATIAL ANALYSIS OF POLY dA.dT TRACTS IN PROXIMAL FLANKING INTERGENIC SEQUENCES

Abundant, long homopolymeric dA.dT tracts were identified within proximal sequences of *Plasmodium spp.* during the Poly analyses. In collaboration with Kenneth Marx's group (University of Massachusetts, Lowell, MA, USA) this work was extended to try and ascertain whether these tracts were apportioned randomly within these regions or whether they exhibited any positional bias. Three Perl scripts were written by Chai-Ho Chen (a PhD student of Kenneth Marx); motif.pl, SB.pl and NN.pl. Motif.pl was a spatial distribution search algorithm whereby the frequency and location of an input 'motif' could be searched for in a specified FASTA sequence and allocated to a user specified bin size. The output data was recorded as a Normalized_count (F_{obs}) whereby the frequency of total matches (number of motifs counted) within a bin was divided by the total bases counted within that bin. SB.pl and NN.pl were scripts to provide controls for the analysis. SB.pl shuffled the DNA input file N-fold times whilst conserving base frequency composition and sequence length. NN.pl was a more complex shuffling script which shuffled the DNA input file as above, but, also conserved the observed nearest neighbour frequencies from the input file within the shuffled output file. Motif.pl was used to identify the observed frequency of non-overlapping homopolymeric tracts within specified upstream and downstream windows of sequence from *Plasmodium spp.* initially - this analysis was then extended to encompass other Apicomplexan parasites. The workflow for Motif.pl is shown in Fig. 2-6 (Chapter 2 Materials and Methods). It should also be noted that NN.pl output is not reported further within this chapter, essentially as it provided little supplementary information to the data provided by the SB.pl shuffling script – it was used here to provide a second set of control data and more widely forms the work of a PhD student in Kenneth Marx's group.

P. falciparum, *P. knowlesi* and *P. vivax* FASTA input files of equivalent length to the Poly work (2000bp upstream and 700bp downstream except in the case of an ORF encounter where the sequence becomes truncated) were utilized for these analyses. Input parameters of N= 5, 10, 15 and 20 homopolymeric dA or dT bases were used as the 'motif' and output data was allocated to 50bp bins. Clear ORF-adjacent poly dA and dT peaks and troughs for all three human malaria parasites were observed (see Fig. 6-9). Interestingly, the length of the ORF-adjacent poly dA or dT tract exhibiting positional bias for each organism did not seem a determining factor as the pattern for N= 5 was echoed by N= 10, 15 and 20 although the frequency of occurrence was considerably reduced for longer homopolymeric tracts. The frequency of these long homopolymeric dA or dT tracts did however, appear to reduce with lower genomic AT-content (see Fig. 6-9 *P. falciparum* and *P. vivax*; 80.6% and 57.7% genomic AT respectively). These data suggested that although homopolymeric tracts are present, in general, at high frequency within the 5' and 3' ORF flanking regions - they appear to have a spatial arrangement ~100bps upstream and ~200bp downstream of the ORF.

In order to refine these observations further, 400bp of sequence centred over the translational start and stop sites were evaluated for the three *Plasmodium* species using N= 5 for poly dA and poly dT and a bin size of 10bps. Fig. 6-10 shows the clear spatial arrangement of homopolymeric tracts adjacent to the ORF for *P. falciparum*, *P. knowlesi* and *P. vivax*. Although the pattern is not identical for all organisms a general demarcation of a poly dT peak ~50bps from and poly dA peak adjacent to the translation start site is observed along with a corresponding poly dA peak adjacent to and a more distal poly dT peak ~100-200bps from the translation stop site. In order to take into account the difference in genomic AT-content of the three organisms a single base shuffle (using SB.pl N= 10 as described above) was undertaken for each of the input data files. Motif.pl was run upon the shuffled files ($F_{10xshuffle}$) files as per the raw data and $F_{obs}/F_{10xshuffle}$ was calculated and these data are shown in Fig. 6-11.

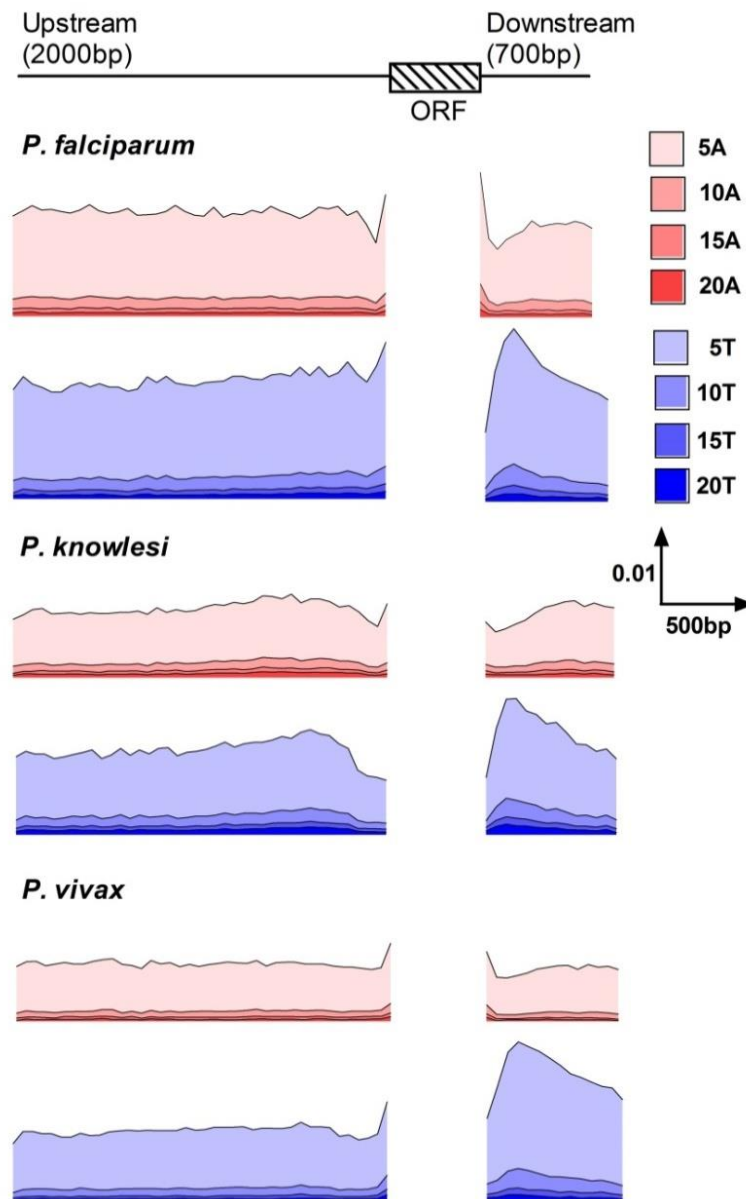


Figure 6-9 Spatial distribution of poly dA.dT tracts in proximal flanking intergenic regions of three *Plasmodium* spp.

The schematics represent the distribution of homopolymeric dA.dT tracts of N=5, 10, 15 and 20 in proximal upstream and downstream flanking regions for three *Plasmodium* spp. Distance from the ORF is plotted against the frequency of each tract (F_{obs}) (bin size = 50bps) and the scale is shown below the key. A similar positional pattern of homopolymeric dA.dT tract distribution adjacent to the ORF is observed regardless of the length of the tract.

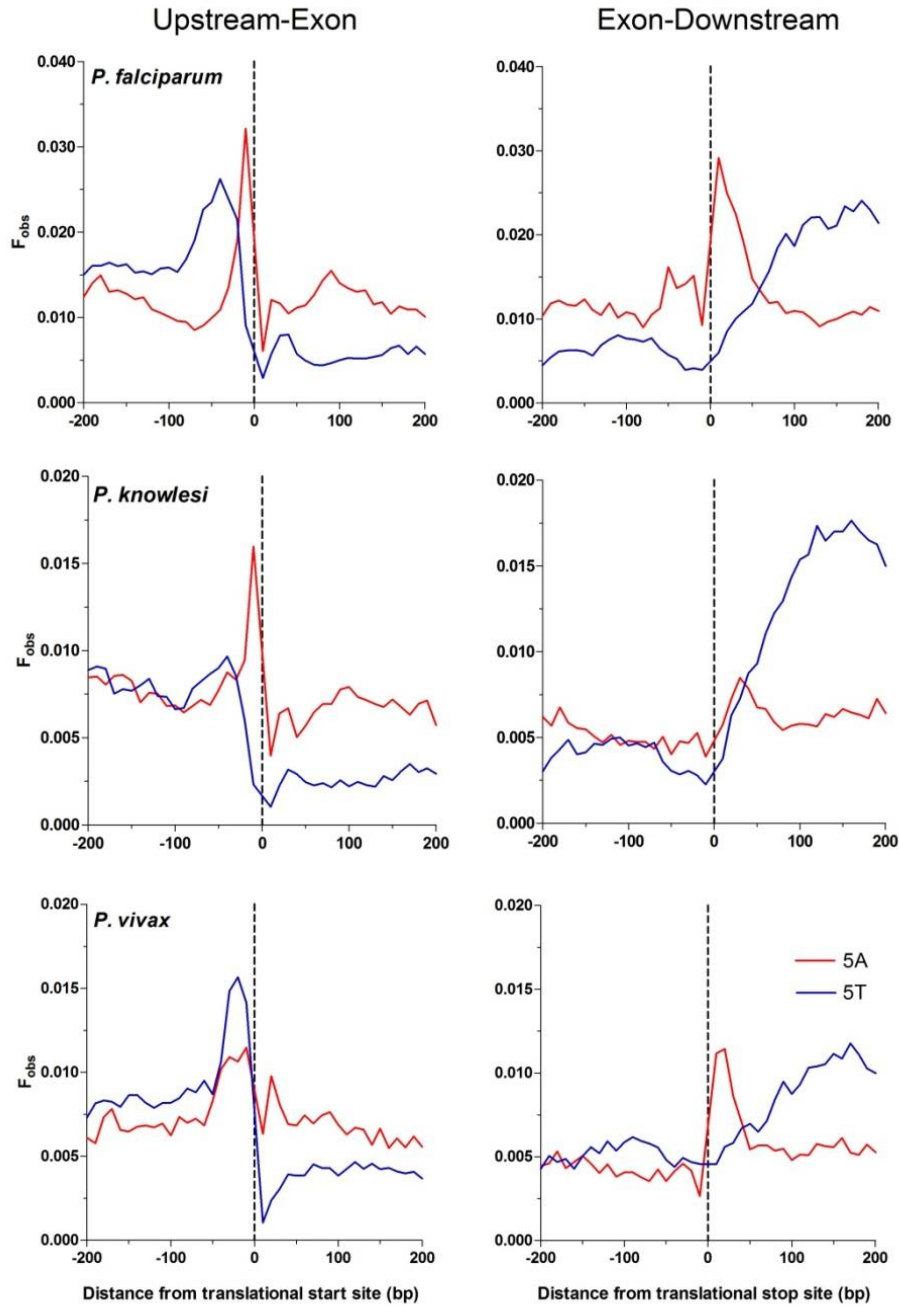


Figure 6-10 Spatial distribution of poly dA.dT tracts over translational start and stop sites of three *Plasmodium spp.*

The schematics demonstrate the spatial arrangement of N=5 poly dA.dT tracts flanking the translational start and stop site for three *Plasmodium spp.* F_{obs} (frequency) (bin size = 10bps) is plotted against distance (400bp flanking the translational start site (upstream-exon) or stop site (exon-downstream)). Red = A and Blue = T. The plots demonstrate the spatial arrangement of poly dA.dT tracts adjacent to ORF boundaries.

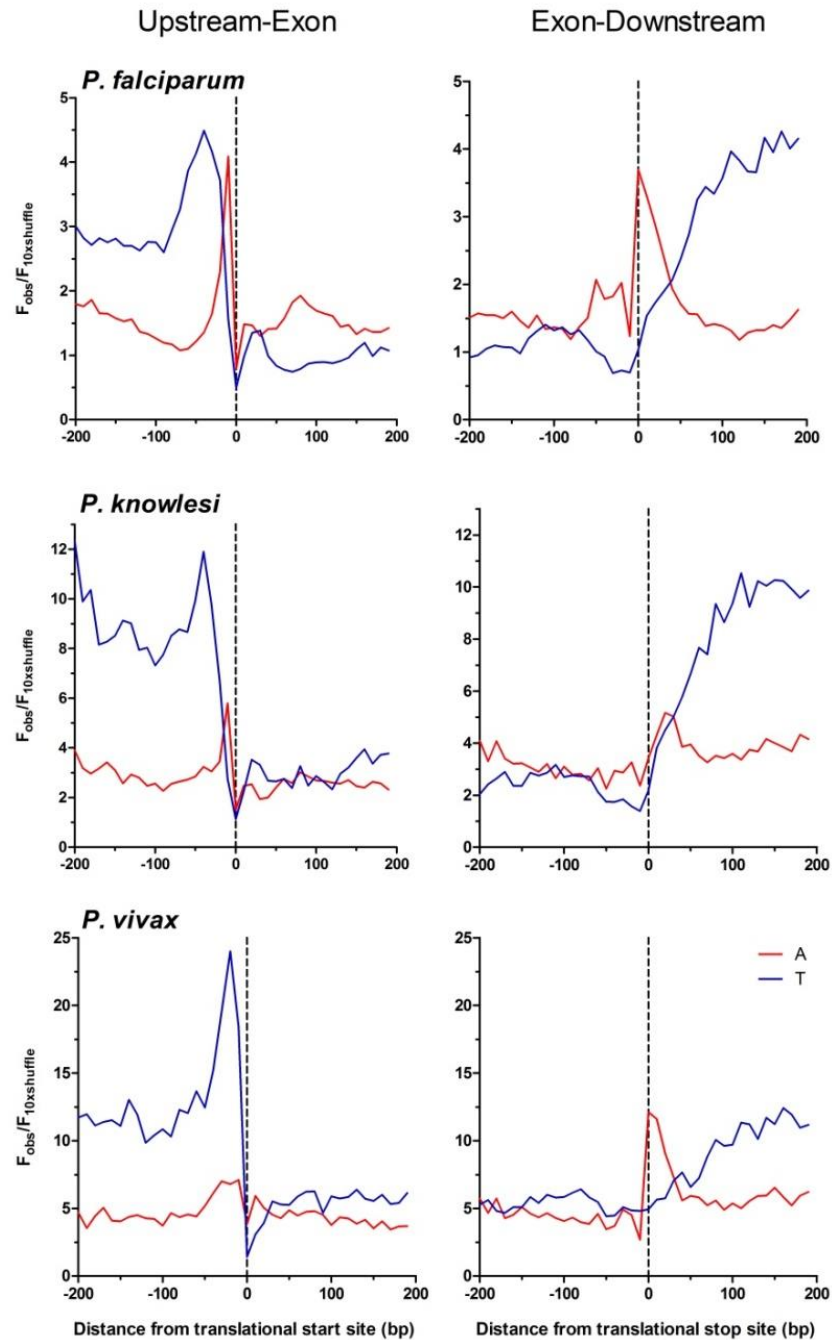


Figure 6-11 Spatial distribution of poly dA:dT tracts over translational start and stop sites of three *Plasmodium* spp. (2)

In order to compensate for the differing genomic base composition in *P. falciparum*, *P. knowlesi* and *P. vivax* a single base shuffle of all input sequences was undertaken ($F_{10xshuffle}$). The ratio of $F_{obs}/F_{10xshuffle}$ was then plotted against distance (400bp flanking the translational start site (upstream-exon) or stop site (exon-downstream) with a bin size = 10bps. Red = A and Blue = T. These plots demonstrate a very similar spatial arrangement of poly dA:dT tracts adjacent to ORF boundaries for all three *Plasmodium* spp.

The use of the shuffle control compensates for the differences in AT-content allowing a more relative comparison to be made between input sequences from organisms of varied AT composition. Therefore, the high F_{obs} of homopolymeric polydA.dT tracts observed in the extremely AT-rich *P. falciparum* genome are rendered comparable to the F_{obs} of the tracts from the lower genomic AT-content *P. knowlesi* and *P. vivax* genomes. The use of the shuffle control demonstrates that, when the underlying base composition is compensated for, that the observed ORF adjacent homopolymeric polydA.dT tract pattern in these three organisms is remarkably similar. These data strongly suggest that this spatial arrangement of poly dA and dT tracts around the ORF is not just a facet of an extreme genomic AT-content but an inherent structural feature surrounding the ORF of human infective *Plasmodium spp.*

In order to evaluate the universality of this spatial arrangement amongst other Apicomplexan parasites, the same 400bp regions were secured for the mouse malaria parasites *P. berghei* and *P. yoelii* along with *Cryptosporidium spp.*, *Theileria spp.*, *T. gondii* and *N. caninum*. Motif.pl was used in the same manner as for the human malaria parasites - where $N=5$ for poly dA and dT with a bin size of 10. What is evident from Fig. 6-12 is that patterns do exist but they do appear specific to different Apicomplexan groups. For example, the translational start site poly dT peak (~50bps upstream of the ORF and clearly prominent in the human infective malaria parasites) does not appear to be anywhere near as prominent in the mouse malaria parasites or *Cryptosporidium spp.* However, the translational start site ORF adjacent poly dA peak does. Staying with the proximal translational start site, the coccidians appear to have a slightly different arrangement which is perhaps more reminiscent of human *Plasmodium spp.* whereas *Theileria spp.*, somewhat surprisingly, (as homopolymeric dA.dT tracts were not found to be over-represented in their genomes) also seem to have a higher 5' ORF adjacent poly dT frequency but no distinct positional peak. However, what is clearly apparent is that the proximal translational stop site arrangement for all these parasites looks remarkably similar with a *c.*200bp downstream region containing an abundance of poly dT.

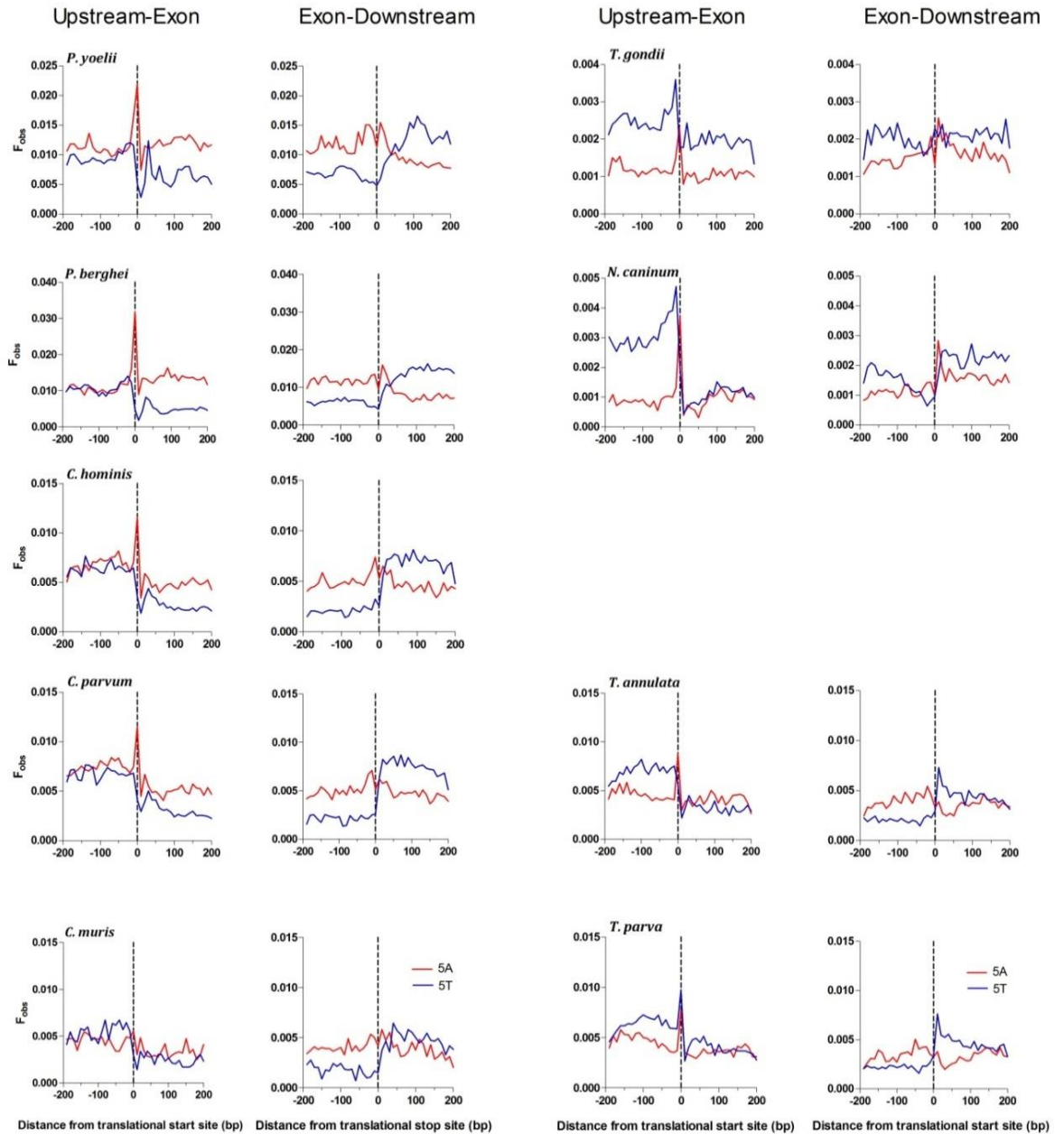


Figure 6-12 Poly dA.dT enrichment proximal to the translational start and stop site of *P. yoelii*, *P. berghei*, *Cryptosporidium spp.*, *Theileria spp.*, *T. gondii* and *N. caninum*

The schematics demonstrate the spatial arrangement of N=5 poly dA.dT tracts flanking the translational start and stop site for other organisms included in this study. F_{obs} , (frequency), is plotted against distance (400bp flanking the translational start site (upstream-exon) or stop site (exon-downstream)). Red = A and Blue = T. Enrichment of poly dA.dT tracts proximal to the translational start and stop sites of the ORF appears to be a feature of intergenic flanking sequences for all organisms used in this study.

As the presence of poly dA.dT has been linked to nucleosome occupancy and genome wide maps of nucleosome occupancy are now available for *P. falciparum* (Ponts *et al.*, 2010) these data were superimposed over the poly dA.dT graphs of the translational start and stop sites (see Fig. 6-13). The nucleosome positioning data presented here (kindly provided by the Le Roch lab) were compiled from two complimentary methodologies, formaldehyde-assisted isolation of regulatory elements (FAIRE) and micrococcal nuclease-assisted isolation of nucleosomal elements (MAINE) (Ponts *et al.*, 2010). Using next generation high throughput sequencing reads from these two methodologies, which identify nucleosome-free (FAIRE) and nucleosome bound (MAINE) DNA respectively, a *P. falciparum* genome-wide nucleosome map was created (Ponts *et al.*, 2010) and these data are presented here as a log₂ ratio. Fig. 6-13 demonstrates that a good correlation exists between the presence of the ORF proximal poly dA.dT tracts and nucleosome positioning; whereby the presence of a nucleosome free region (NFR) coincides with the occurrence of homopolymeric poly dA.dT tracts (where N≥5) and nucleosomes are positioned over the 5' and 3' end of the ORF which is less densely populated with poly dA or dT. These data suggest that the spatial arrangement of homopolymeric dA.dT tracts around ORF in *P. falciparum* could represent intrinsic determinants of a nucleosome positioning code in this organism.

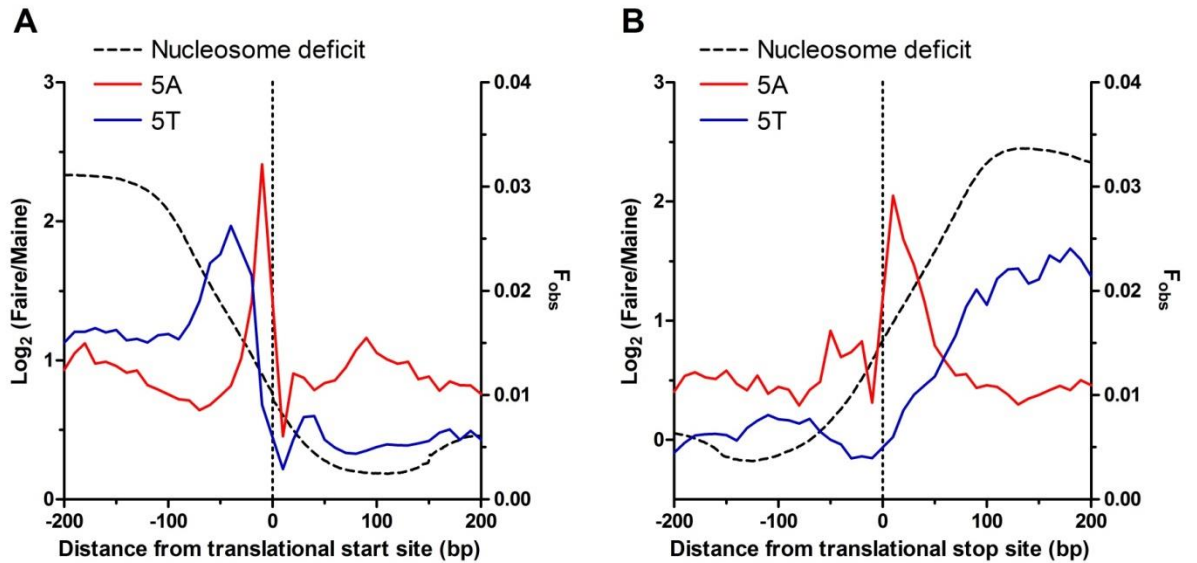


Figure 6-13 Spatial distribution of nucleosome occupancy and poly dA.dT tracts over translational start and stop sites of *P. falciparum*

The schematics demonstrate the spatial arrangement of N=5 poly dA and dT tracts flanking the translational start and stop site for three *Plasmodium spp.* Red = A and Blue = T. F_{obs} (frequency) (bin size = 10bps) is plotted against distance (400bp flanking the translational start site (upstream-exon) or stop site (exon-downstream)). Overlaid over these data is the Ponts *et al.*, 2010 \log_2 (Faire/Maine) nucleosome occupancy data (shown in the form of a dotted line) whereby peaks represent nucleosome deficiency and troughs represent nucleosome occupancy.

6.4 SPATIAL ANALYSIS OF POLY dA.dT TRACTS OVER THE CORE PROMOTER IN *P. FALCIPARUM*

Owing to the paucity of mapped TSSs in *P. falciparum*, a comparison of the homopolymeric dA.dT arrangement around TSSs was not possible here. However, using predicted core promoters with EGASP scores of 0.4-1.0 (where 1.0 denotes highest confidence) (Brick *et al.*, 2008) this analysis could be attempted on a region previously characterised by Ponts *et al.* (2011) for nucleosome positioning data. A clear pattern of homopolymeric dA (Fig. 6-14B) and dT (Fig. 6-14C) tract organisation is observed – not surprisingly the F_{obs} for these peaks decreases with lower EGASP scores as more, less confidently predicted core promoters are

included in the analysis. In addition, the presence of these homopolymeric dA.dT tracts correlates well with Log_2 FAIRE/MAINE nucleosomal occupancy data available (Fig. 6-14A) (Ponts *et al.*, 2011). By comparing the 3477 EGASP scores of highest confidence (1.0) for both poly dA and dT of N= 5 and N= 10 and overlaying this with the FAIRE/MAINE nucleosome positioning output a robust evaluation is attained (see Fig. 6-15). A clear peak of poly dT is observed ~10-20bps upstream of the predicted TSS in conjunction with a less abundant peak of poly dA 30~50bp further upstream of the poly dT. Previous work has demonstrated that although the homopolymeric tract length may vary, it is the positioning that appears more invariant, therefore, both N= 5 and N = 10 were evaluated and indeed both demonstrate the same pattern albeit with the N = 10 having a lower F_{obs} . The overlay of the nucleosome positioning data again correlates a NFR with the presence of the homopolymeric tracts suggesting again that these two facets are likely not coincidental. It should be noted, however, that the EGASP scores used for this analysis were taken from a dataset of bioinformatically predicted promoters - the algorithm for which was trained using physiochemical predictors - one of which being the structural effect of DNA bending, so some caution needs to be applied in the interpretation of these results (Brick *et al.*, 2008).

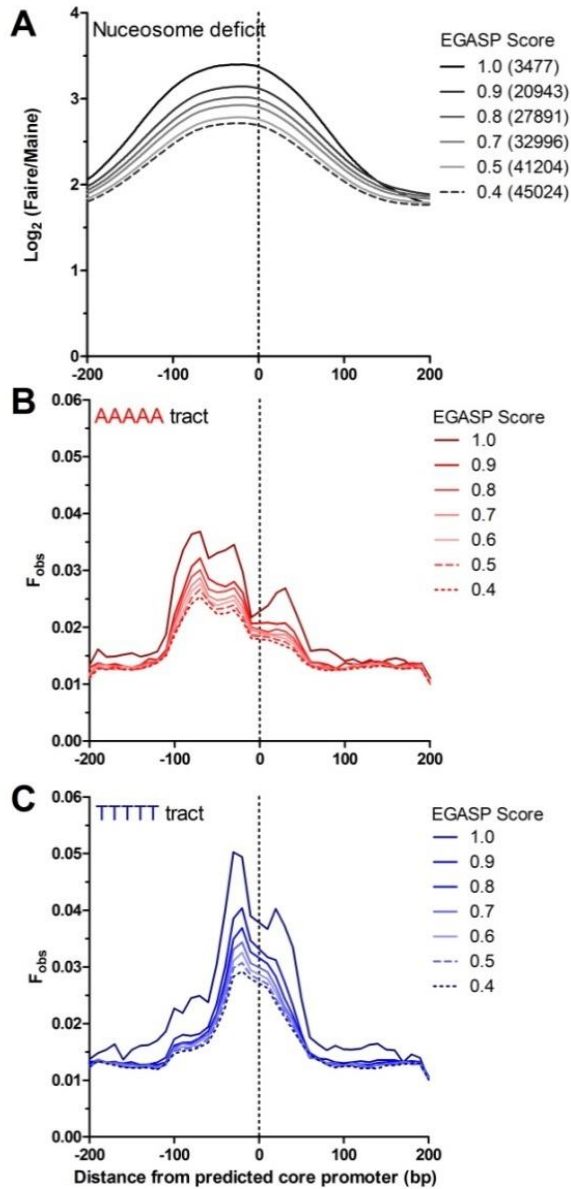


Figure 6-14 Spatial distribution of nucleosome occupancy and poly dA.dT tracts over *P. falciparum* predicted core promoters of varying confidence

The Malarial Promoter Predictor (MAPP) tool was used to predict core promoter regions (Brick *et al.*, 2008). Scores were derived from positive predictive value and sensitivity (using EGASP criteria) where 1 is highly confident whereas 0.4 is only moderately confident. Using these data both (A) nucleosome positioning data (\log_2 FAIRE/MAINE sequence reads, Ponts *et al.*, 2011) and F_{obs} of N=5 poly dA (B) and F_{obs} of N=5 poly dT (C) were plotted over a 400 bp region centred over the predicted core promoter. The number of predicted core promoters is shown in (A) in parenthesis and the EGASP scores are shown in the key. Red = A and Blue = T. Nucleosome deficiency is denoted by a dotted line where the increase in FAIRE/MAINE ratio indicates nucleosome deficient regions.

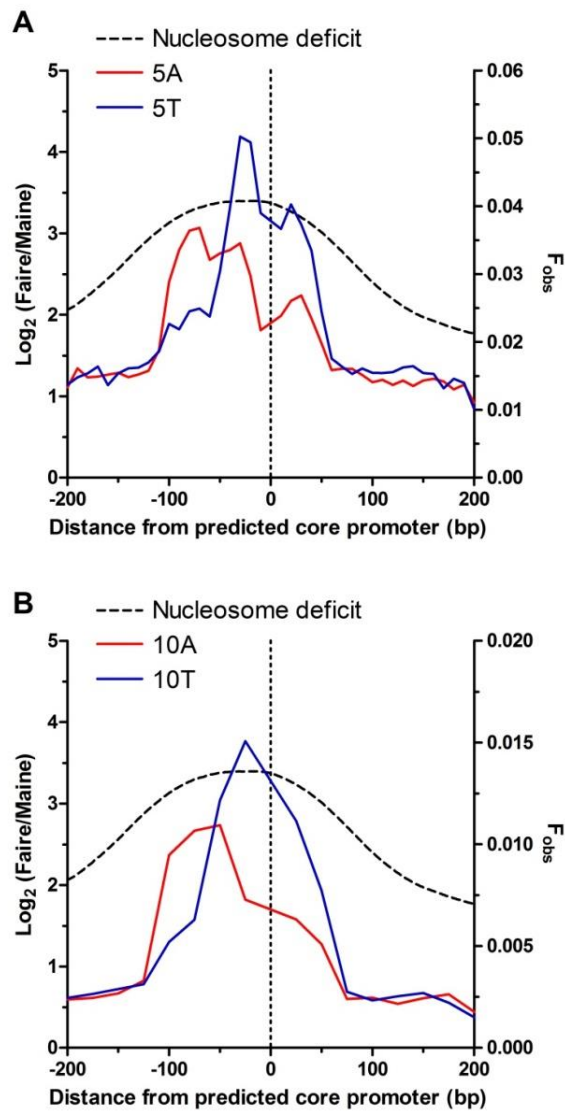


Figure 6-15 Spatial distribution of nucleosome occupancy and poly dA.dT tracts over predicted core promoters in *P. falciparum*

The plot of spatial distribution of observed frequency shows only data with an EGASP score of 1.0 (Brick *et al.*, 2008) along with the nucleosome positioning data (\log_2 FAIRE/MAINE sequence reads, Ponts *et al.*, 2011). The F_{obs} of (A) $N=5$ and (B) $N=10$ (10 base bins) for homopolymeric dA and dT tracts is plotted over a 400 bp region centred over the predicted core promoter. Red = A and Blue = T. Nucleosome deficiency is denoted by a dotted line where the increase in FAIRE/MAINE ratio indicates nucleosome deficient regions.

6.5 DISCUSSION

The Apicomplexan parasites investigated within this study demonstrate three distinct organisational groups in terms of the overproportionment (P) and overrepresentation (R) of homopolymeric tracts within their gene flanking regions. Group 1 (*Plasmodium spp.* and *Cryptosporidium spp.*) contains over-abundant short homopolymeric dG.dC and long homopolymeric dA.dT tracts. Group 2 (the coccidians) contains over-abundant long homopolymeric dG.dC and short homopolymeric dA.dT tracts whereas Group 3 (the piroplasmida) contains no over-abundance of homopolymeric tracts at all. There is also a second clear demarcation between all other parasites investigated here and the Plasmodiidae, whereby *Plasmodium spp.* appear unique in containing an overabundance of short homopolymeric dG.dC and long homopolymeric dA.dT tracts within their CDS as well. The organisation of homopolymeric tracts does not appear to be random, but instead demonstrates clear patterns of spatial distribution around the coding sequences. For *P. falciparum*, a similar enrichment of poly dA.dT is also observed adjacent to highly predicted core promoters. Critically, the spatial arrangement of homopolymeric dA.dT tracts around the ORF and putative TSS correlates with previously mapped NFRs suggesting poly dA.dT tracts may operate as intrinsic regulators of nucleosome positioning in *P. falciparum*.

Comparative analyses of the ORF-flanking regions for a wider group of Apicomplexan parasites, demonstrates that: i) the degree of overrepresentation of homopolymeric tracts is unrelated to the underlying base composition however, ii) the threshold of overrepresentation (here defined where $R > 0.5$) is related to the underlying base content and not a single simple thermodynamic threshold of around 7-10bp. These findings are in accordance with Zhou *et al.* (2004), although their study represented an analysis of a diverse range of eukaryotes – as opposed to the more focussed study presented here (Zhou *et al.*, 2004). In addition, a third and novel finding is reported here - that the degree of tract length

overproportionment, within compact genomes at least, appears unrelated to base composition, but, instead correlates directly with the length of the IGR itself. Further exploration of this observation will require analysis of other eukaryotes that display features of compact genome organisation – such as yeasts and fungi.

The origins and function of homopolymeric dA.dT tracts have been a matter for debate for decades. Two leading theories have been proposed for their origin; 1) slipped-strand replication errors and 2) seeding from polyadenylated transcripts of retrotransposons. Slipped-strand replication errors are thought to occur in homopolymeric dA or dT tracts when they reach the thermodynamic threshold of $N \geq 7$ whereupon the length of the less stable poly dA or dT tract increases logarithmically during replication (Dechering *et al.*, 1998). Data presented here regarding the base-composition effect on thresholds for tract expansion and the preferential spatial organisation of tracts does not support such a model in Apicomplexan parasites. The effect of selective pressure resulting in some tracts being more liable to slipped-strand expansion than others could account for the preferential spatial arrangement we observe, but does require a more complex hypothesis to explain how tracts are seeded and expanded. Insertion of homopolymeric tracts by transposable elements, found ubiquitously in metazoan eukaryotic genomes, is therefore an alternative hypothesis (Zhou *et al.*, 2004). However, to date, no evidence exists for the presence of transposable elements in modern Apicomplexan genomes (DeBarry and Kissinger, 2011) but it is plausible that they were present in more ancient lineages (Roy and Penny, 2007). The absence of over-abundant homopolymeric tracts in Piroplasmida (although ORF adjacent poly dA.dT tracts were evident) somewhat complicates this hypothesis as Piroplasmida share common origins within the sub-order Haemosporidia and thus the Plasmodiidae family - where tracts appear most over-abundant. However, the rapidly evolving genomes of Apicomplexan organisms in general, encompassing dramatic genomic arrangements, gene and intron loss and species

specific gene expansion, suggests that transposable element activity within ancestral lineages should not necessarily be precluded (Roy and Penny, 2007).

The spatial arrangement of homopolymeric dA.dT tracts around the ORF strongly suggests that these tracts may impart functionality. The observed correlation with nucleosome positioning data suggests that, as hypothesized in other systems, these poly dA.dT tracts could act as DNA sequence codes for nucleosome positioning. However, most research in this area has concentrated on the positioning of the -1 and +1 nucleosomes at the TSS in an open (or TATA-less) promoter whereby a wide NFR is created possibly to allow transcription factor access (Wu and Li, 2010). TSS, for most of the organisms investigated to date, are close to the start of the ORF. For *P. falciparum* this seems unlikely, as the ORF adjacent poly dA.dT tracts observed would be unlikely to be incorporated into the TSS as previous chapters in this thesis suggest that most transcription start sites for *P. falciparum* are likely in excess of 600bps upstream of the ORF. However, nucleosome mapping data does suggest the presence of a NFR adjacent to the ORF. The physical barrier model proposed by Mavrigh *et al.* suggests that the strong positioning of some nucleosomes, e.g. the +1 and -1 nucleosomes, could physically restrict and therefore dictate the placement of neighbouring nucleosomes, thus, via this statistical positioning effect a nucleosomal stacking system could be created which loses integrity over distance from the primary nucleosome placement (Mavrigh *et al.*, 2008). In order for this model to work it would be necessary, on a genome wide scale, to have a certain percentage of strongly positioned nucleosomes. It is plausible therefore, that the positions of these 'core' nucleosomes (though not all) could be denoted by the underlying DNA sequence perhaps in the form of strategically placed homopolymeric dA.dT tracts (Korber, 2012). The positions of the poly dA.dT tracts - adjacent to the translational start and stop site could, for example, in *P. falciparum* whose chromosome internal regions (with the exception of the centromeres) reside in a relatively euchromatic state, demark nucleosome stacking over the ORF. The ORF 5' preferential accumulation of ancestral histones after replication and the

deposition of new nucleosomes within the ORF is discussed by Korber in his review (Korber, 2012). However, debate on the relative contribution of the underlying DNA sequence to nucleosome positioning continues and it seems unlikely that it acts alone - more likely in concert with energy dependent trans-acting factors and remodelling enzymes such as SWI/SNF and ISWI and RSC (remodels the structure of chromatin) reviewed by (Korber, 2012) and (Arya *et al.*, 2010). The apparent preference of the double histone variants H2A.Z and the Apicomplexan specific H2B.Z for poly dA.dT sequence combined with the stronger acidic surface patch on H2A.Z - which could promote internucleosomal interactions, suggests that histone variants and their myriad of potential modifications may also have roles to play (Hoeijmakers *et al.*, 2013; Arya *et al.*, 2010). CpG methylation has also been implicated in the alteration of nucleosome DNA preferences and *P. falciparum* genome-wide maps of asymmetric cytosine methylation (Me⁵C) have recently been published by Ponts *et al.*, (Ponts *et al.*, 2013). Therefore, nucleosome positioning *in vivo* is likely to be a thermodynamic combination of all or many of these factors, but these data, in conjunction with other research, suggest that poly dA.dT may have a role to play (Segal and Widom, 2009a; Segal and Widom, 2009c).

Interestingly, the strong poly dT tract positioning directly upstream of the putative TSS in *P. falciparum* is similar in arrangement to that observed in *D. discoideum*, (a similarly AT-rich organism), where it was hypothesized that abortive transcription and RNA pausing - thought to be a key facet of the cascade of gene expression required for multicellular development (Levine, 2011), could result in the production of a less-stable poly-rU RNA/DNA hybrid (Chang *et al.*, 2012). It is interesting to note, within this context, that the constitutive pre-assembly of RNA Pol II PIC components has been reported for *P. falciparum* (Gopalakrishnan *et al.*, 2009) throughout the IE cycle suggesting that perhaps *P. falciparum* RNA Pol II may also reside in a 'poised' state in order to accommodate the increased processivity demands

placed upon the parasite during the trophozoite stage of the IE life-cycle (Bozdech *et al.*, 2003; Sims *et al.*, 2009).

The presence of long over-represented homopolymeric tracts in *Plasmodium spp.* CDS is unique to this family of Apicomplexan parasites. Codon bias is known to exist within *Plasmodium spp.* but it is different for the more AT-rich malaria parasites - these preferring poly-lysine and poly- asparagine (predominantly coded for by AAA and AAT respectively) whereas *P. vivax* and *P. knowlesi* with less AT-rich genomes have a preference for poly-alanine (coded for by GCN) (Dalby, 2009). Furthermore, the third codon preference for A and T in highly expressed genes is only prominent in the more AT-rich *Plasmodium spp.* and is therefore more likely a product of underlying base composition (Yadav and Swati, 2012). Taking these data together it seems unlikely that codon bias alone could account for the over-abundance of long poly dA.dT homopolymeric tracts observed within the CDS of *Plasmodium spp.*. However, codon bias likely explains the high Base-A (v Base-T) content identified within *P. falciparum* CDS (evident as an over-representation of poly dT). It is also known that parasitic host-changing organisms, such as the Apicomplexa, who have complex life-styles with different hosts and environmental changes to accommodate generally have much more disordered genomes (Pancsa and Tompa, 2012). Low complexity regions, comprising runs of a single amino acid repeat sequence, are well documented in the genomes of *Plasmodium spp.* with at least one insertion present in up to 90% of *P. falciparum* genes (Frugier *et al.*, 2010). The presence of these LCRs accounts for the difference in genome size between *Plasmodium* and yeast - whereby the quantity of genes are roughly comparably but the size of the *Plasmodium* protein, in general, is much larger (Frugier *et al.*, 2010). *P. falciparum* LCRs have recently been sub-grouped into three types; i) a heterogenous group (87% AT-rich) with reduced complexity but few recognizable amino acid repeat patterns ii) a high heterozygosity and high AT-content group containing many long asparagine degenerate repeats (coded for by AAT and AAC) and iii) a GC-rich group containing multiple indels - suggestive of a function

in recombination (Zilversmit *et al.*, 2010). These data are interesting, as the heterogenous group of LCRs could perhaps account for the homopolymeric poly dA.dT detected in this analysis. Unfortunately, whilst many different roles have been hypothesized for LCRs in general such as: host-evasive immunodominant regions, a protective response to heat shock and, interestingly, control of translation efficiency, no specific function has currently been assigned to this group (Frugier *et al.*, 2010; Zilversmit *et al.*, 2010).

In summary, this chapter provides a comprehensive comparative analysis of homopolymeric tract abundance, length and positional bias for a cohort of Apicomplexan parasites. The over-representation and spatial arrangement of long homopolymeric dA.dT tracts in the genomes of *Plasmodium spp.* are particularly consistent - regardless of genomic AT-composition, and suggest that these tracts may impart functionality. Whether the presence of the ORF adjacent homopolymeric dA.dT tracts influences processes such as transcription or translation will be evaluated in Chapter 7.

The data presented in Chapter 6 has just been accepted for publication and is attached as Appendix H

Russell K, Cheng C, Bizarro JW, Ponts N, Emes RD, Le Roch K, Marx KA, Horrocks P. Homopolymer tract organization in the human malarial parasite *Plasmodium falciparum* and related Apicomplexan parasites. *BMC Genomics* 2014, 15:848

CHAPTER 7 FUNCTIONAL ANALYSIS OF THE 5' UTR IN *P.*

FALCIPARUM

7.1 INTRODUCTION

To date, most studies of promoter function have concentrated on mapping transcriptional start and stop sites and/or deleting/modifying sequences residing outside of the UTR - such as those immediately upstream of the transcription start sites, to explore their role in directing the control of core promoter activity - see summary Table 1-1 for list of studies (Horrocks *et al.*, 2009). A smaller number of studies have been carried out (Horrocks and Kilbey, 1996; Porter, 2002; Militello *et al.*, 2004; Brancucci *et al.*, 2012) that specifically explore responses to deletions that would be expected to occur within the 5' UTR. These studies tend to indicate a profound reduction in absolute reporter gene expression upon the reduction in size of the 5' UTR. However, some of these studies have been marred by the use of unmatched 5' and 3' sequence and/or the use of transient transfection techniques. Given these limitations, interpretation of findings can be a challenge as such approaches may be subject to artefacts such as the loss of plasmid and/or the incorrect assembly of chromatin over the reporter plasmids (Horrocks and Lanzer, 1999). Furthermore, none of these studies, to date, have explored how the deletion of 5' UTR affects the temporal expression profile of the gene under investigation.

Work reported in this thesis (Chapter 6) presents evidence for a preferential positional bias of poly dA.dT tracts in proximal flanking sequences immediately adjacent to the ORF. These homopolymeric tracts would undoubtedly reside within the UTR and thus have the potential to act as functional elements in the control of gene expression. Previous studies in *P. falciparum* suggest that the deletion of homopolymeric dA and/or dT tracts affects absolute levels of transcription (Porter, 2002; Polson and Blackman, 2005). However, a systematic

evaluation of the role such tracts may play in transcription and/or translation, or, an evaluation of the plasticity or rigidity of the UTR length within these context, has, to date, not been carried out.

This chapter presents an analysis of the effect of a series of deletions of putative 5' UTR regions for a gene that shows a strong temporal pattern of trophozoite expression. Here I report the use of integration methodology, to ensure stable selection of a reporter construct, and the use of matched 5' and 3' flanking sequences to investigate the effect that the removal of sequences located within the 5' UTR has upon the absolute level and temporal expression profile of the reporter gene.

Declaration: This work was carried out in collaboration with Sandra Hasenkamp and has been published (Hasenkamp *et al.*, 2013. Functional analysis of the 5' untranslated region of the phosphoglutamase 2 transcript of *Plasmodium falciparum*. *Acta Tropica* 127; 69-74) attached as Appendix B. Here, I will describe my work on the generation of the transfection constructs used in this study and the subsequent reporter gene assays. Additional work on the confirmation of the transcription start site and quantitative analysis of stage-specific mRNA was completed after I had left the laboratory and is only presented here in the discussion (Appendices E3-5).

7.2 SELECTION AND CHARACTERIZATION OF PFD0660W AS CANDIDATE GENE FOR THIS STUDY

Several criteria were important when selecting an appropriate candidate gene for this study. Specifically: (i) it needed to be a highly expressed gene with a clear pattern of stage specific-expression, (ii) preferably with EST data available to help delineate the necessary UTR flanking regions, (iii) it needed a reasonable sized 5' and 3' IGR, to assist in ready PCR amplification, lacking certain key restriction sites necessary for the sub-cloning strategy and

(iv) the immediate 5' and 3' flanking regions needed to have the characteristic spatial arrangement of poly dA.dT tracts.

PF0660w (re-annotated on PlasmoDB to PF3D7_0413500) is located on *P. falciparum* chromosome 4 (Fig. 7-1A) and codes for the phosphoglutomutase-2 protein which is a component of the glycolysis pathway, reversibly catalysing the conversion of 2-phosphoglycerate to 3-phosphoglycerate (Hills *et al.*, 2011). This essential protein represents a typical housekeeping gene expressed during the IE cycle, with maximal transcript steady-state levels during the trophozoite stage (Fig. 7-1B). It has 1286bps of 5' flanking upstream sequence and 674bps of 3' flanking downstream sequence. These regions lack *ApaI*, *PstI*, *HindIII* and *SalI* restriction sites - necessary for the planned sub-cloning strategy. There was EST data available on dbEST (though not on FullMalaria) and RNASeq data was also available (although not outside of the ORF). It also contained the spatial organisation of poly dA.dT tracts of interest (see Appendix E-1). Therefore, this gene appeared a good match for the proposed experimental work.

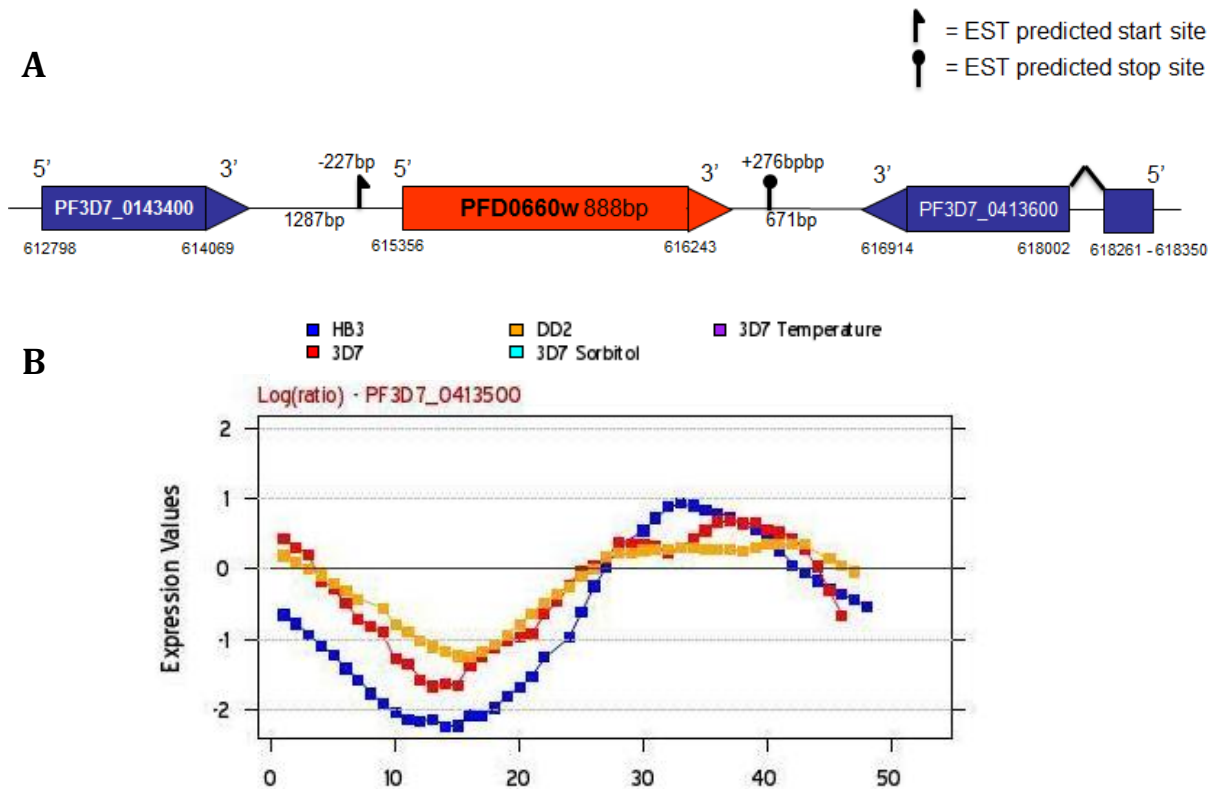


Figure 7-1 PFD0660w overview and expression profile

PFD0660w (now re-annotated to PF3D7_0413500) is located on chromosome 4 and resides between PF3D7_0143400, a *var* exon1 pseudogene and PF3D7_0413600 a putative 26S proteasome AAA-ATPase subunit RPT3 (PlasmoDB). EST data (dbEST) predicts a transcription start site -227bp upstream and a transcription stop site +276bp downstream of the ORF, respectively. B Available microarray data (PlasmoDB) shows a strong trophozoite-stage IE expression profile.

7.3 GENERATION OF REPORTER CONSTRUCTS FOR THIS STUDY

The subcloning strategy employed is shown in Fig. 7-2. The approach adopted was to replace the sections of an existing reporter cassette (pmPLP1) designed by Ellie Wong (a previous PhD student of the Horrocks laboratory) in a study of the expression of the *Pfpcna* gene (Wong *et al.*, 2011). The entire reporter cassette in pmPLP1 was flanked by *ApaI* and *PstI* restriction sites. The *luc* gene within the reporter construct was flanked by two *HindIII* restriction sites. In order to replace the 3' downstream (DS) region the *luc* and 3' *Pfpcna* were excised using *HindIII* and *PstI* and the remaining backbone purified. A PCR product was

produced of the entire PFD0660w 3' DS region (3D7 gDNA used as template) with proximal *HindIII* and *Sall* restriction sites and a distal *PstI* restriction site added (See Fig. 7-3 for schematic of oligonucleotide positions, Table 2-2 for oligonucleotide sequences and Appendix E-2 for oligonucleotide positions within DS sequence). The sequence integrity of this and all other PCR products generated here was confirmed using a commercial service (Eurofins Mwg, Germany). This DS sequence was then ligated into the remaining pmPLP1 backbone. Four different PCR products were then produced for the 5' upstream regions (US), (3D7 gDNA used as template). The first was a full-length (FL) PFD0660w 5' US construct (1286bp), then a series of three PFD0660w 5' US truncated products of 1083, 872 and 607bps respectively (See Fig. 7-3 for the relative oligonucleotide positions compared to that of the FL construct, Table 2-2 for oligonucleotide sequences and Appendix E-2 for oligonucleotide positions within US sequence). Each of these products had a distal *ApaI* restriction site and a proximal *HindIII* site. The cloning plasmid was then restricted with *ApaI* and *HindIII* to allow the ligation of each 5' US product respectively. A further PCR product of the *luc* gene (amplified from the pmPLP1 plasmid), with a 5' *HindIII* site and a 3' *Sall* restriction site added, was then inserted after cutting the cloning plasmid with *HindIII* and *Sall*. This strategy provided for the synthesis of four different constructs: 1) FL with full-length 5' PFD0660w US, a *luc* reporter gene and a full-length PFD0600w 3' DS, 2) $\Delta 1$ with a 1083bp 5' PFD0660w US, a *luc* reporter gene and a full-length PFD0660w 3' DS, 3) $\Delta 2$ with a 872bp 5' PFD0660w US, a *luc* reporter gene and a full-length 3' PFD0660w DS and 4) $\Delta 3$ with a 607bp 5' PFD0660w US, a *luc* reporter gene and a full-length 3' PFD0660w DS. The *ApaI/PstI* reporter cassette from each of these plasmids was then inserted into an existing pDCAttP plasmid called Delta 1 (Nkrumah *et al.*, 2006; Wong *et al.*, 2011) to generate four plasmid constructs in a transfection-ready plasmid that contained two critical additional components; (i) a blasticidin drug selection cassette and (ii) the *attB* sites necessary for integrase mediated

insertion into an *attP* locus located on chromosome 7 of the *P. falciparum* parasites to be genetically modified (Nkrumah *et al.*, 2006).

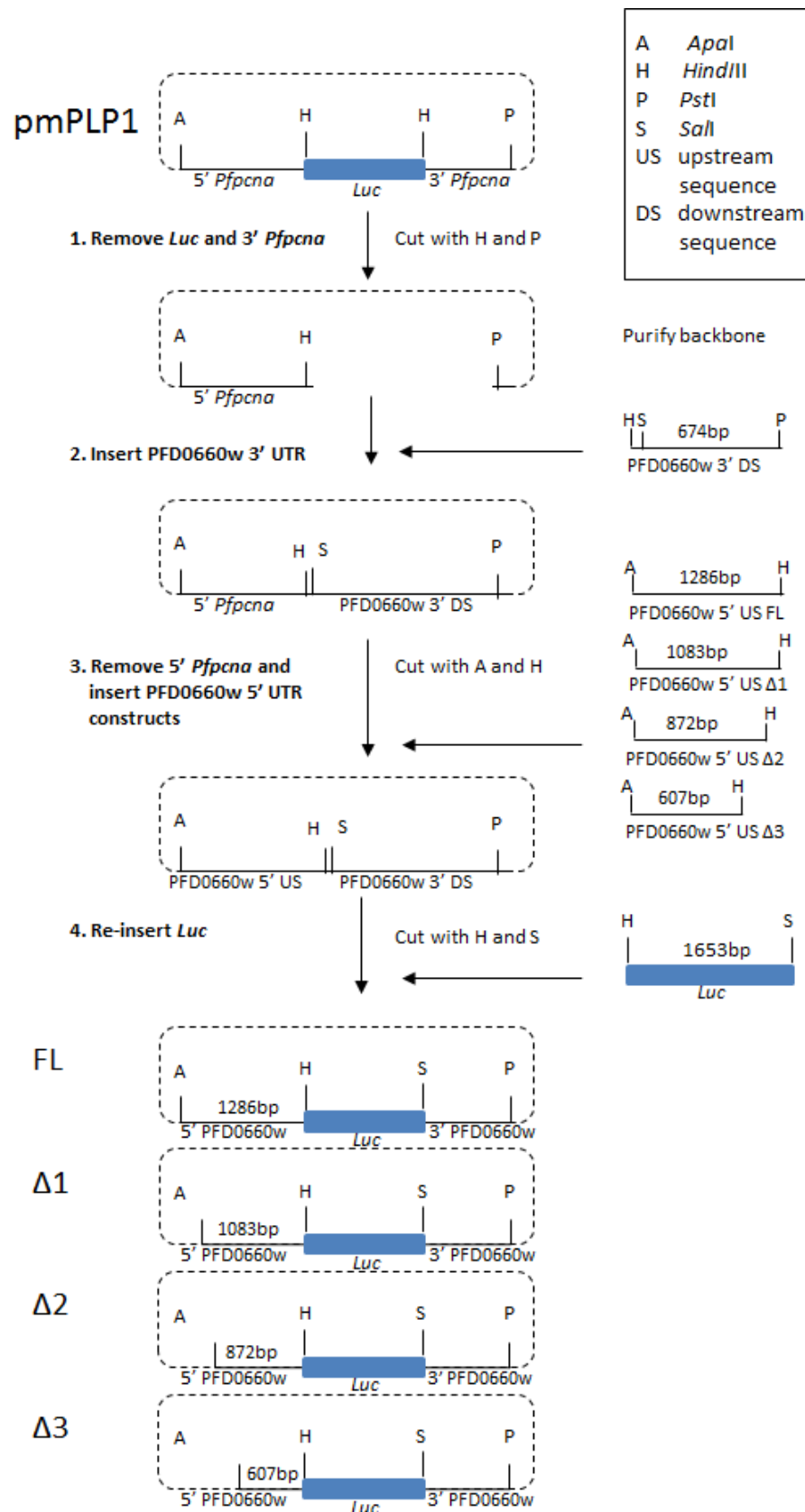


Figure 7-2 Schematic representing the sub-cloning strategy employed here

See main text for details.

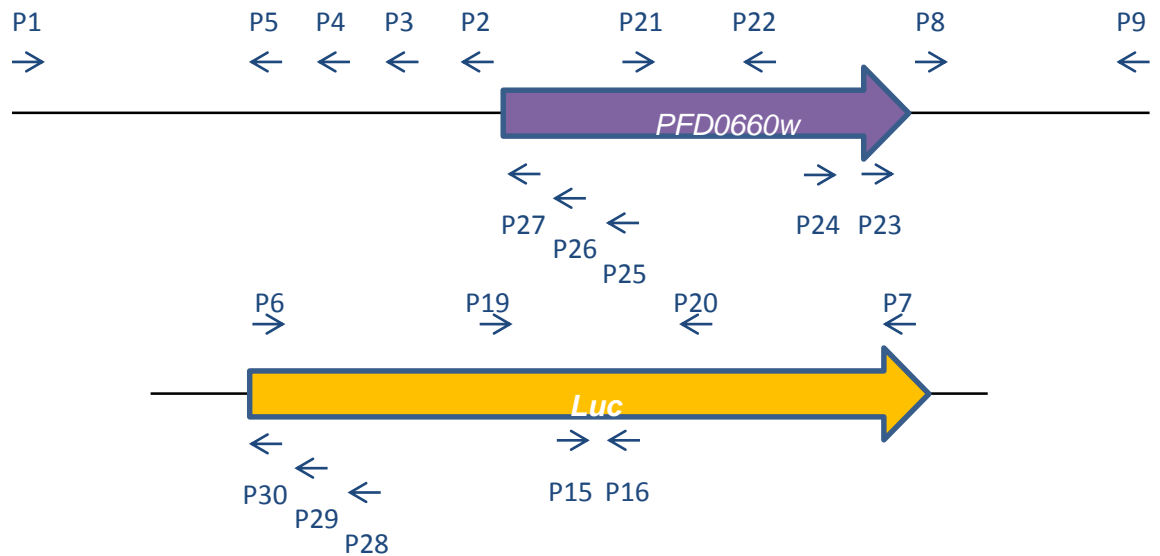


Figure 7-3 Primer locations over *luc* and PFD0660w ORF

Schematic demonstrating the location of all primers used during this study over the *luc* gene, the PFD0660w gene and its 5' and 3' upstream and downstream sequences. All primer sequences and corresponding numbers are listed in Tables 2-1, 2-2, 2-3, 2-4 and 2-5 (Materials and Methods).

In order to confirm the size of the final plasmids to be used for transfection (FL, $\Delta 1$, $\Delta 2$ and $\Delta 3$), each was subjected to restriction digest analysis with: 1) *ApaI/HindIII*, 2) *ApaI/PstI*, 3) *ApaI/NotI* and 4) *ApaI/SacI* respectively (Fig. 7-4). This analysis confirmed the sizes and relative position of all components of the reporter constructs. The plasmid restriction maps for each of the four constructs (FL, $\Delta 1$, $\Delta 2$ and $\Delta 3$) are shown in Fig. 7-5 along with a schematic depiction of the upstream flanking sequence present in each construct (Fig. 7-6).

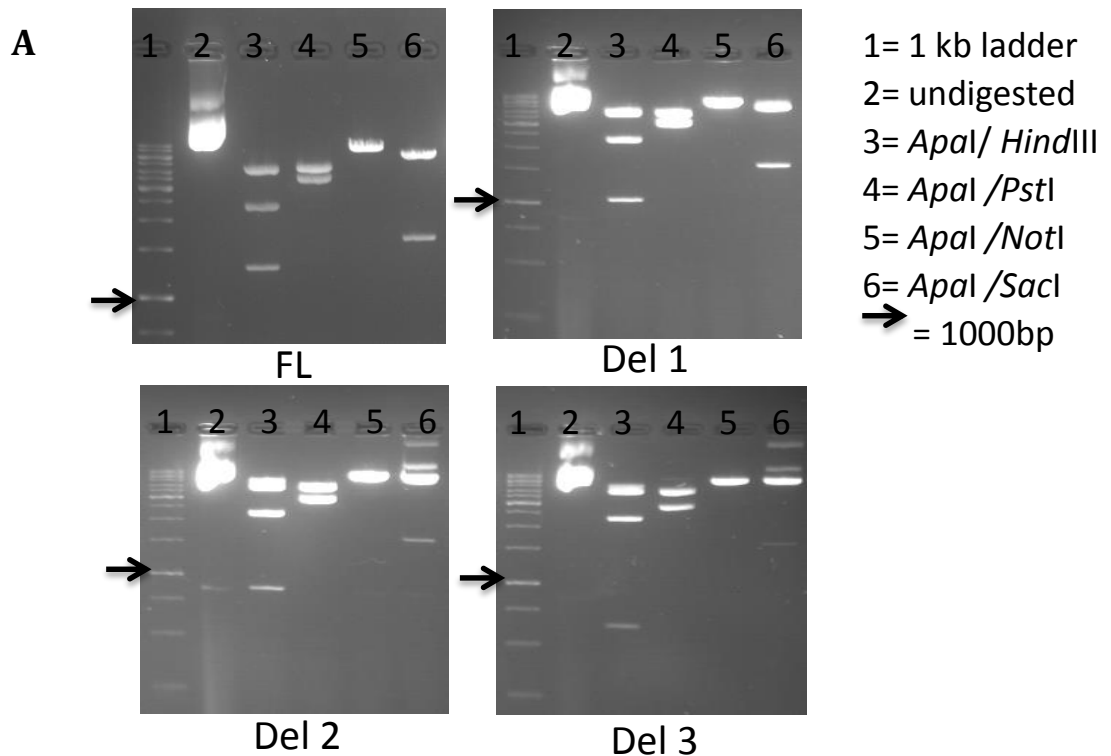
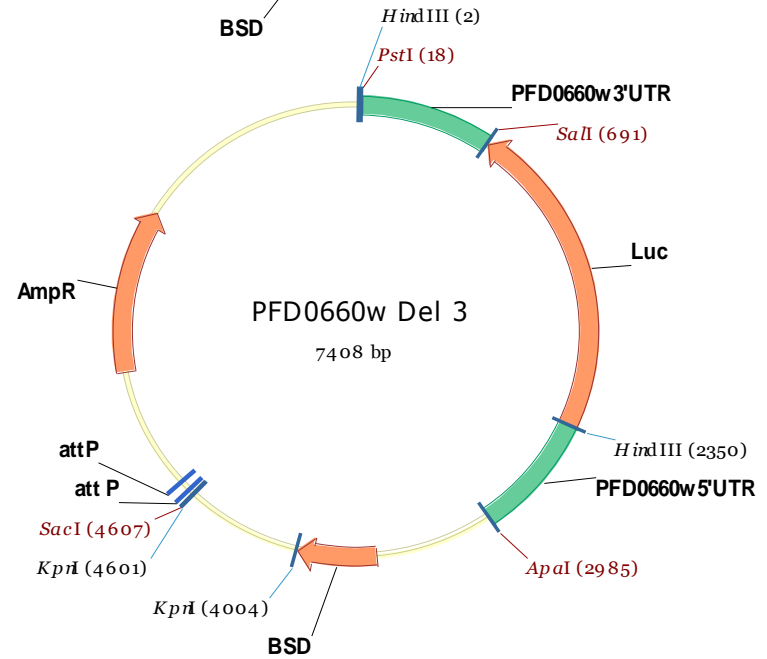
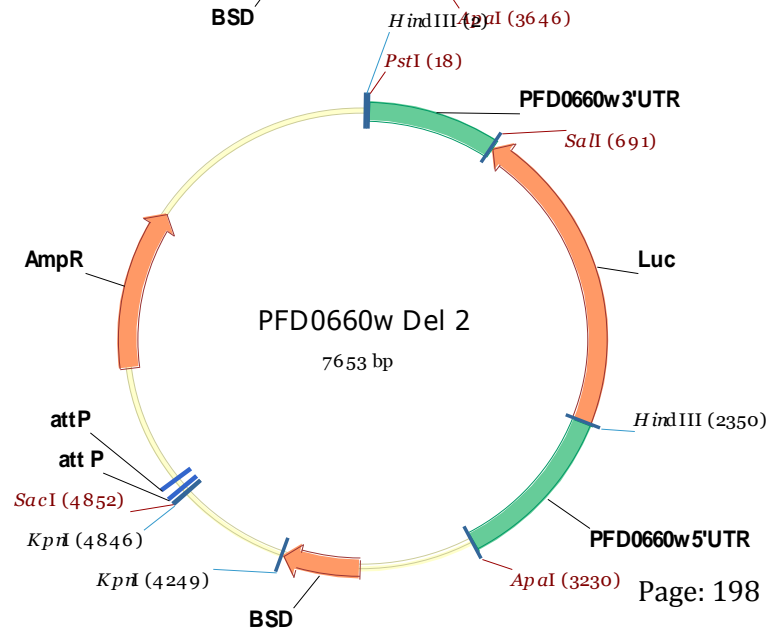
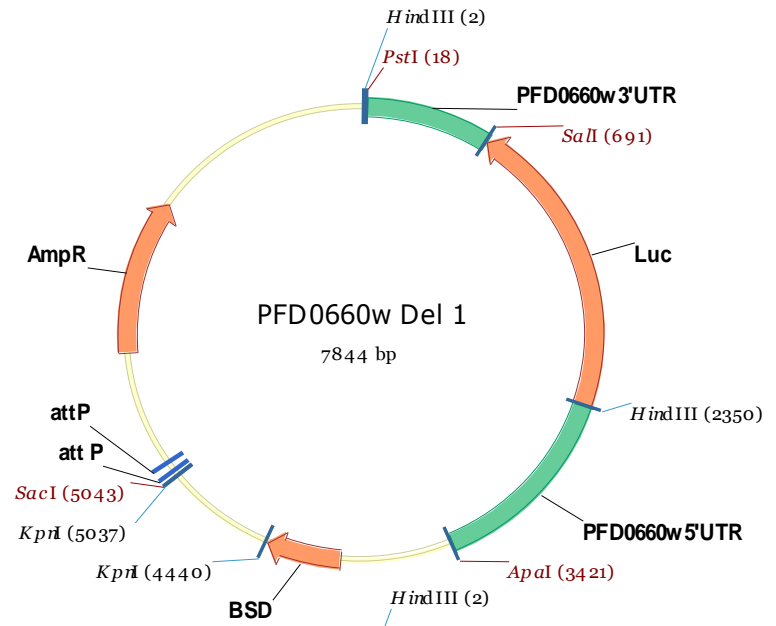
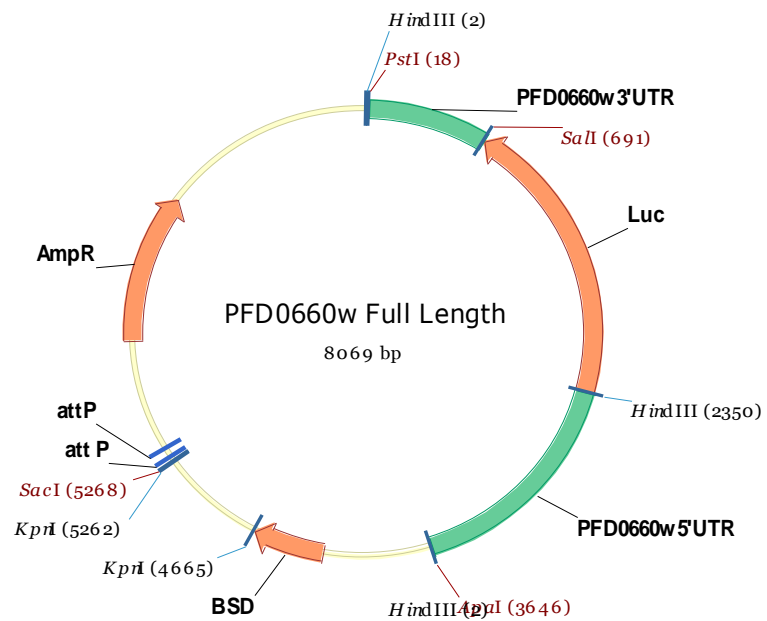


Figure 7-4 Confirmation of reporter plasmids by restriction digest

Reports the results of restriction digest analysis of constructs FL, $\Delta 1$, $\Delta 2$ and $\Delta 3$ following size-fractionation gel electrophoresis (post-stained with ethidium bromide). Lane 3 demonstrates the expected reduction in size of the 5' upstream sequence from the FL construct down to $\Delta 3$. Lane 4 demonstrates the expected reduction in size of the whole reporter cassette from the FL down to $\Delta 3$. *NotI* was present in the *Pfpcna* 5' upstream sequence and as such only uncut plasmid is detected in Lane 5. *SacI* resides behind the two *AttP* integration sites and thus this restriction enzyme in conjunction with *ApaI* removes only the BSD gene. Lane 6 demonstrates the presence of the expected larger and smaller band for all constructs. Note Lane 6, $\Delta 2$ and $\Delta 3$, also contains two additional larger fragments of undigested plasmid - likely in supercoiled and open supercoiled forms

Figure 7-5 Restriction plasmid maps

(overpage) For each of the constructs FL, $\Delta 1$, $\Delta 2$ and $\Delta 3$ a restriction plasmid map is shown illustrating the relative position of the luciferase reporter cassette to the blasticidin S deaminase (BSD) drug resistance cassette and the two *attP* integrase sites.



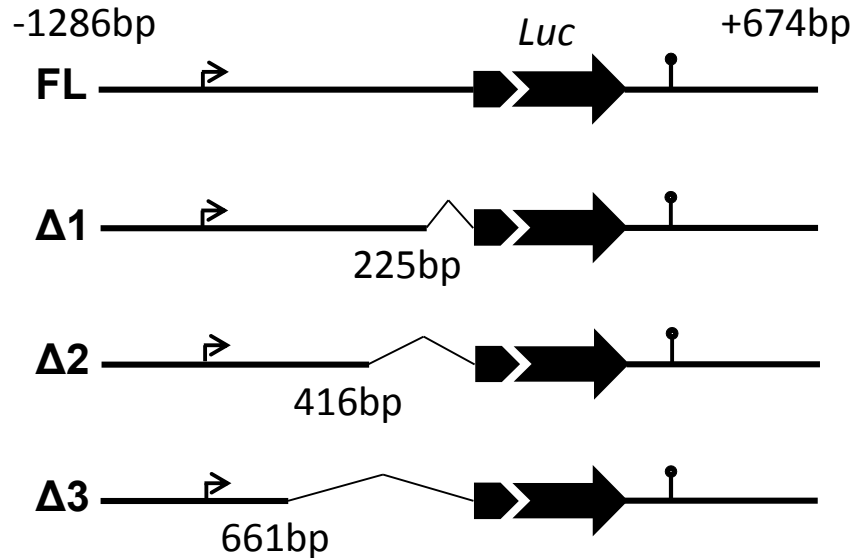


Figure 7-6 Schematic depiction of the position of upstream flanking sequences in FL, Δ1, Δ2 and Δ3 constructs used in this study.

The length of sequence under each of Δ1, Δ2 and Δ3 constructs indicates the length of 5' proximal flanking sequence deleted in each case.

7.4 TRANSFECTION OF FL, Δ1, Δ2 AND Δ3

All constructs were transfected as per the 'direct' transfection protocol (Hasenkamp *et al.*, 2012a). In brief, 40μg of pINT (contains the integrase enzyme and is maintained by neomycin drug resistance marker for G418) and 40μg of the pDCAttP plasmid were transfected into intraerythrocytic ring stage *Dd2^{AttB} P. falciparum* culture via electroporation. After 48hrs recovery, blasticidin S hydrochloride (2.5μg/ml) (Invitrogen, UK) and G418 sulphate (100μg/ml) (Sigma, UK) was added to the culture media. A schematic representation of the anticipated integrase-mediated insertion of the luciferase reporter cassette into the *cg6* locus on chromosome 7 is shown in Fig. 7-7.

Recovery time varied between 22 and 34 days post-transfection and was detected upon observation of recrudescence drug-resistant parasites on Giemsa-stained thin smears. Upon

parasite detection, pINT was cured via the withdrawal of G418. Blasticidin S hydrochloride was also removed from the culture medium for two weeks and then re-applied to cure the culture of unintegrated plasmid. Two independent clones were created for the transfectant series (FL-Δ3). Whilst both were confirmed positive for luciferase expression, only the first series were taken forward for further study here.

The direct transfection protocol utilized for this work has been published (Hasenkamp, S., Russell, K.T., Horrocks, P. Comparisons of the absolute and relative efficiencies of electroporation-based transfection protocols for *Plasmodium falciparum*. 2012 *Malar J.* 21;11:210) doi: 10.1186/1475-2875-11-42 **and is attached as Appendix G**

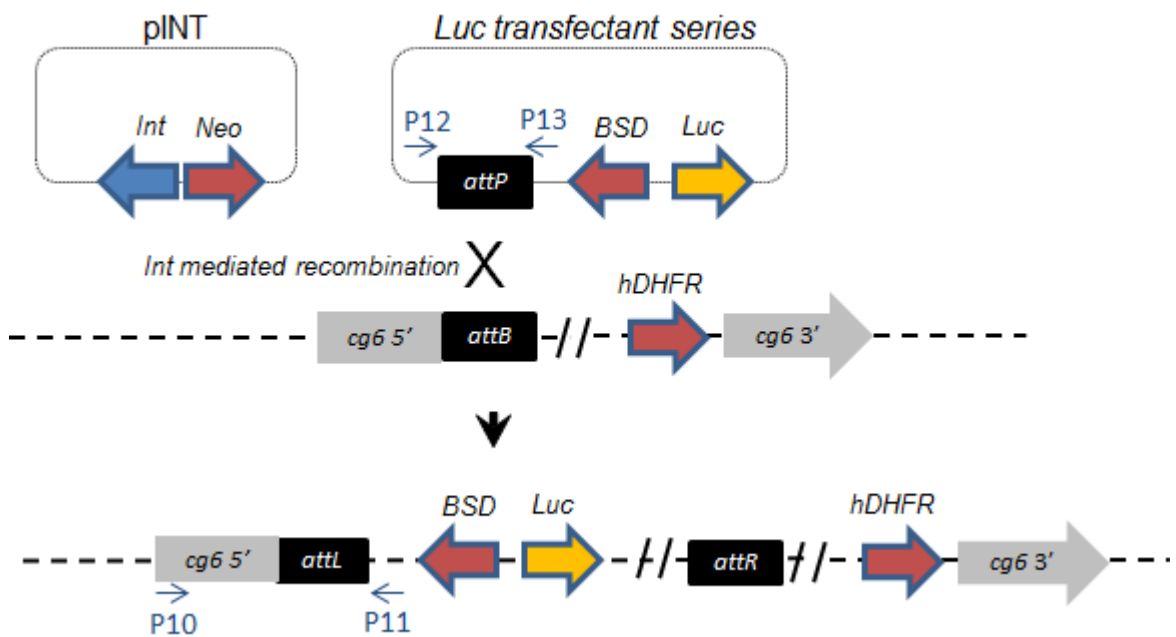


Figure 7-7 Schematic representation of integrase-mediated insertion of the luciferase reporter cassette into the *cg6* locus on chromosome 7

The *luc* transfectant series and the pINT plasmid were introduced into *P. falciparum* (Dd2^{attB}) culture via electroporation. The *attP* site on the transfectant plasmid recombines, in the presence of the integrase (provided by the pINT cotransfected plasmid), with the *attB* site within the *cg6* gene on chromosome 7 of the Dd2 parasite. This process produces *attL* and *attR* sites and hence the irreversible integration of the transfection cassette into a known location within the parasite genome (Nkrumah *et al.*, 2006).

P. falciparum Dd2^{attb} contains an *attB* insertion site within the *cg6* (glutaredoxin-like) gene on chromosome 7 disruption of which has been determined to be non-detrimental for asexual blood stage growth (Nkrumah *et al.*, 2006). Upon plasmid integration the genomic *AttB* site within the *cg6* gene is replaced with an *AttL* and *AttR* site (see Fig. 7-7). Therefore, PCR amplification was carried out over the *AttL* junction to confirm integration (primer sequences Table 2-3). Unfortunately, genomic integration could only be demonstrated for the FL and $\Delta 1$ transfectants (see Fig. 7-8A). The absence of integration for all plasmid constructs is in line with previous lab experience. However, PCR amplification over *AttP* (primer sequences Table 2-3) demonstrated the presence of stably maintained episomal plasmid in all four cultures (see Fig. 7-8A) despite drug-cycling. Once again, drug cycling does not always remove episomal plasmid and by ceasing drug cycling at this stage the loss of $\Delta 2$ and $\Delta 3$ transfectant lines was prevented.

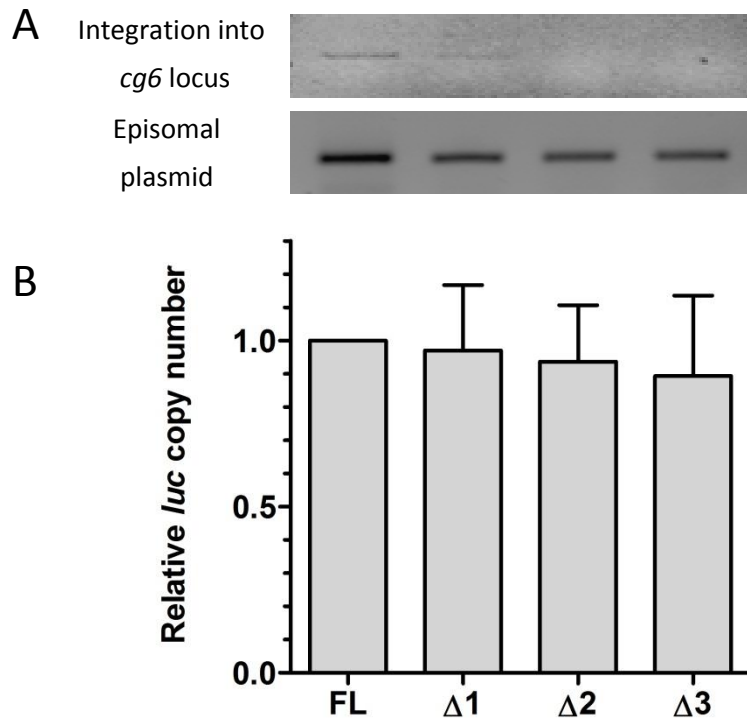


Figure 7-8 Confirmation of Integration and qPCR

(A) Genomic integration was demonstrated for the FL construct and Δ1 only. However, episomal plasmid was present within all four of the transfectant series (FL, Δ1, Δ2 and Δ3). (B) qPCR demonstrated that there was no significant difference in the relative *luc* copy number for FL, Δ1, Δ2 or Δ3 (mean normalized *luc* count (to seryl-tRNA synthetase) relative to the FL transfectant \pm StDev (n=3, P>0.05 ANOVA)).

It was now important to ascertain the relative plasmid copy number so quantitative PCR (qPCR) was carried out over the *luc* gene in all four transfectant lines (primers sequences Table 2-4). Three independent experiments were carried out and all *luc* comparisons were made against the FL transfectant (mean normalized *luc* count to seryl-tRNA synthetase) relative to the FL transfectant \pm StDev (n=3, P>0.05 ANOVA)). Using the $2^{-\Delta\Delta Ct}$ method of relative quantitation it was established that there was no significant difference between relative copy number in any of the transfectant series (FL, Δ1, Δ2 or Δ3) (see Fig. 7-8B). Whilst unable to produce genetically-modified parasites with solely genomically integrated reporter constructs, the evidence that similar amounts of reporter cassettes were present in each transfectant allowed the investigation to progress.

7.5 ANALYSIS OF LUCIFERASE EXPRESSION IN THE FL TRANSFECTANT

To establish the temporal profile of luciferase expression in the FL transfectants, RNA was harvested from 3 morphologically distinguishable IE time-points; ring (12-18hrs), trophozoite (18-28hrs) and schizont (28-38hrs) stages using the Trizol (Invitrogen) method of isolation (Kyes *et al.*, 2000) (see Chapter 2 Materials and Methods for the full procedure). The RNA (5µg) was size fractionated on a TBE gel, then Northern Blotted onto Hybond N+ membrane which was then hybridised with α -³²P labelled PFD0660w and *luc* probes (oligonucleotide sequences shown in Table 2-1). The size of the transcript for PFD0660w was 2kb (gene size 888bp giving a UTR of ~1.1kb) and the size of the *luc* transcript was 2.7kb (gene size 1.6kb giving a UTR of ~1.1kb) (Fig. 7-9). The northern blot data sizes were in accordance with expectations and the correct temporal profile for *luc* transcripts was observed (during the latter stages of IE development) which suggested that the elements necessary to reconstitute the correct temporal control from endogenous transcription start and stop sites were present in the FL construct.

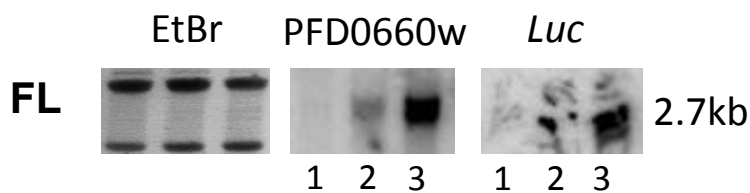


Figure 7-9 Stage-specific northern Blot analysis of the FL construct

RNA was harvested from well synchronized *P. falciparum* culture at 3 morphologically distinguishable IE time-points 1) ring stage, 2) trophozoite stage and 3) schizont stage. The RNA was run on a 1% TBE gel and transferred via Northern Blot. The probes used (shown in Table 2-1 materials and methods) were specific to either PFD0660w or to *luc*. Both the PFD0660w and *luc* blots demonstrated temporal expression profiles consistent with existing expression data; no expression during ring stage, expression increasing at trophozoite stage and high expression during the schizont stage. The size of the PFD0660w transcript was 2kb. The size of the *luc* transcript was 2.7kb which was consistent with the increase in reporter gene size.

EST data for PFD0660w (<http://www.ncbi.nlm.nih.gov/dbEST>) identified putative transcription start and stop sites at -227bp and +276bp. These data would indicate that the PFD0660w transcript to be in the region of 1391bps – considerably smaller than the *c.* 2Kb shown by northern blot, reflecting another example of variance in these two methods. In order to try and qualify this disparity 5' and 3' rapid amplification of cDNA ends (RACE) was undertaken using RNA from mature trophozoites (see Table 2-5 for oligonucleotide primers). Two 3' RACE clones identified a transcription stop site at around +276bps downstream of PFD0660w and *luc*, which was consistent with the EST data. Moreover, a putative polyadenylation signal was also apparent at this position upon inspection of the downstream flanking sequence (Fig. 7-10). It is worth noting at this time that a 3' UTR of 276 bases for a 1.1kb UTR suggests a 5':3' apportionment ratio of almost exactly 75:25. If the northern blot data was accurate, this would predict a transcription start site in the region of 825bp upstream of the ORF. Repeated attempts at 5' RACE, using oligonucleotide primers from both PFD0660w and *luc*, resulted in a range of short RACE products being produced - none extending further than -246bp upstream of the ORF. These data were comparable with the EST data. However, RACE can be problematic within such AT-rich sequences and loss of processivity can occur. Analysis of the sequences in the region of the RACE and EST clones reveals an extensive 27-mer polyT tract (Fig. 7-10). Therefore, a 5' RACE was also carried out on the *luc* transcript from the $\Delta 3$ transfectant - this already had 661bp of 5' flanking sequence in the UTR deleted and it was hoped that this may overcome the poor processivity issues over this long polydT tract. Using this approach, a RACE product of ~200bp was consistently generated. Unfortunately, all attempts at cloning this product failed. The inability to produce a sequenceable 5' RACE product from $\Delta 3$ prompted further investigation in the form of another deletion construct, $\Delta 4$, by the laboratory after the completion of the experimental work for my PhD. $\Delta 4$ had 943bp of upstream sequence deleted, which included the region of

the predicted TSS at or around -825bp, and no transcription was revealed from this cassette (referred to later within my discussion with evidence presented in Appendix E-3).

```
Upstream Sequence
>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2) | genomic |
Pf3D7_04_v3 forward|(geneStart-1286 to geneStart+0) | length=1287
ACCATACACAAATAAACATAGACACAGATACCTCCACCCACAATACATATATAAATAGCA
ACAAACAAACTAATACTTTCTTACCAATTTAACAAGTCTACATATAACAATATCTAGAT
ACACACACATATCTCTCAATCTAAATAAATGGCATGTAAAATACACCAACATAACTATGA
TATATTATATTATATAAATAAATTTAAATGTTATGTTATATATGTAAAGATAAAAAATCT
TATATATAAAGAGAATAGTAAAATTTAAAATATAAAAATGTGATTTTATTTTTTAAATA
TGTTTATATTTGAAAGGGGAGAAAAGATATAAAAATAAATGAGAAAATAATTATTTAAATA
AAAAATTTATATGATTTTTAATAAAAATTTTTATTATATATACATATAAATTTTGAT
TGTTATACATTTCTCTTTTGATATATATTTTTATAAGTGTTTGTGTTAAATTTAAAT
ATAAATAAATAAATTTGTGGTTAAAAATAAAAAATGATAATATGAAATAATATATATA
TATATATATTTATAAATAAATTAAGATAAAAATAAATTTCTAAGGATTTTAGTATATTTTTG
GTTTTCACTATAATTTAATGATATATATATATATATATATATATTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTCATATGTTGAGTAAAAAATATGTATGTGAATACACAGAACCA
TAAAAATAATTTTTAAGAATAATAAATAAATGAATGAATTTTACTTTTTTAAATAT
ATTTTTGTATTTGTAAATAAATAATTATTATTATTGATTATATATATATATATTTTT
TTTTTGGTGAGGATTTAGTGC GTTAAATATTTTGCATATTTTCAATATAAATTTGAAATAA
TATAATACATATATATATATATTTTTTTTTTTTTTTTTTTTTATATTTTAAATTTAAT
TTTTAAATTTTATTATTATATATATTTTTAAAAATTATATTTCTTTAAAAATATTATTTAT
TTTTAAATTTTTATGTTCCAGATGTGTTATTTTTACACATTTTTTTTTATTGATTTTTT
TTTTTTTTTTTTTTTTTTTATGTAATAAATTGTTGTAATTATTATTAATTTGGATTTT
CATAAATAATACATATATTACAAAATATATATATATATATATATATATATATATATAT
ATGATATATTTTTATTATTTTTTTTTTTTTTTTTTTTTGTTAAAATAAAAAAGTGTGTGAGAATATA
TATAAAAAAAAAAAAAATATGAGTAAAAAA

Downstream Sequence
>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2) | genomic |
Pf3D7_04_v3 forward|(geneEnd+0 to geneEnd+674) | length=675
ATAATTCCTTTGAATTATGAAAAAAAAAAAAATATATATTTACATAGAATATATAAATC
TATTAATATGATAATATGGTAAATATATATATATATAAATATATATATATATATATAT
GTATATATGTGTATGCTATTTTTTTTTTTTTTTTTTTTTATATATTTTTTAAAAACACTA
GATAATTTATTTAATTATATCTCTTATATATTTTTTTTTTAAATATTATTTAAAAATACA
TAATAAATTTATTTAACGGAATTATAAAAAAAAAATTTATATTGTTTATATATATATATAT
ATATATATATATATATATATATTACGTTAAAAGAAAGAAAAAAGAAAAAAGATAATA
TTGTTGATAGTTAATATTAGATGGATGTAATTTTTGTATAAATCCATATAAATATTTTAC
TTTTCAATATAAAAAAAAAAAAAAAAAAAAAAAAAAATTTGATTTTAAACAATATAAATATCTATA
TAATAATAATAAAAAAAAAAAAAAGATGCCACAATAAAAAATATCAGTTGAGCACATTTTA
TAGTCATATAATATGTGTGTTTGAACCAATCCTGTGTCATTAAAAATAAATATAATACAAT
ATATATATATATATATACATATATGTATTTTTTTTATTATATTCTGTAAATCAT
TATAACATTCGTTAT
```

Figure 7-10 Predicted UTR from dbEST

Flanking upstream and downstream sequence for the PFD0660w ORF with dbEST predicted UTR in red. The red circle denotes the presence of a long homopolymeric dT tract located at the position of the dbEST predicted transcriptional start site.

7.6 ANALYSIS OF ABSOLUTE AND TEMPORAL EXPRESSION IN THE DELETION CONSTRUCT SERIES

To explore the effect of deletion on transcript size and temporal transcription, stage specific northern blots were undertaken for $\Delta 1$, $\Delta 2$ and $\Delta 3$. No change in the temporal expression profile was observed for any of the deletion constructs - all appeared to exhibit trophozoite/schizont stage expression (Fig. 7-11). Therefore, the timing of gene expression did not seem to be affected by the deletion of up to 661bps of 5' ORF adjacent upstream sequence. A transcript was produced for all of the deletion series which suggested that all constructs retained the same putative transcription start site (TSS) - located in the region of 825bp upstream of the ORF. In addition, the size of the transcript reduced in each of the transfectants $\Delta 1-3$ by approximately 200bps, in accordance with the expected reduction in size of the 5' flanking sequences cloned. Unfortunately, northern blot data does not represent a complete quantitative analysis therefore any reduction in expression level could not be accurately ascertained here. However, this aspect was also addressed by the laboratory after I had completed my experimental work and quantitative RT-PCR was carried out on early to mid trophozoite and mature trophozoite/schizont cDNA to ascertain relative *luc* transcript copy number and this work will again be commented upon in the discussion to this chapter (Appendix E-4).

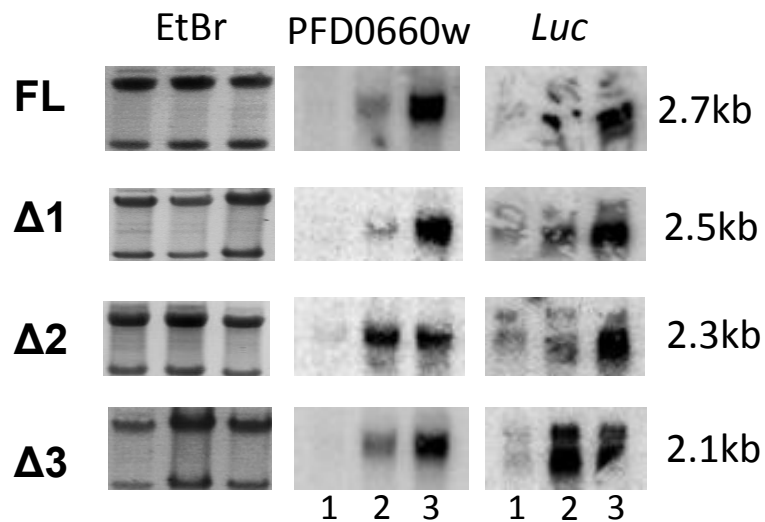


Figure 7-11 Northern Blots of the transfection series

RNA was harvested from well synchronized *P. falciparum* culture for each of transfectant series (FL, Δ1, Δ2 and Δ3) at 3 morphologically distinguishable IE time-points 1) ring stage, 2) trophozoite stage and 3) schizont stage. The RNA was run on a 1% TBE gel and transferred via Northern Blot. The probes (shown in Table 2-2 materials and methods) were specific to either PFD0660w or to *luc*. Both the PFD0660w and *luc* blots demonstrated temporal expression profiles consistent with existing expression data; no expression during ring stage, expression increasing at trophozoite stage and high expression during the schizont stage. Reductions in band size were also observed for the *luc* gene and these were consistent with the reduction in the length of the 5' UTR in the transfectant series. Intensity comparisons of PFD0660w did not identify any discrepancy in *luc* transcript signal between the FL construct or any of the deletion constructs.

Luciferase reporter assays were then undertaken according to the single-step lysis protocol (Hasenkamp *et al.*, 2012b) developed by this laboratory. Prior to starting experimental procedures, all cultures (FL, Δ1, Δ2 and Δ3) were diluted to a 1% parasitaemia whilst maintaining a 2% haematocrit. 200µl samples were taken throughout the growth cycles at 5 timepoints; i) T1 early rings (6-12hpi), ii) T2 late rings (12-18hpi), iii) T3 early trophozoites (18-26hpi), iv) T4 mature trophozoites (26-34hpi) and v) T5 schizonts (36-42hpi). 40µl samples from each 200µl aliquot were used, in triplicate, for the luciferase assay and three independent experiments were carried out (ie. n=9). The remainder of the aliquot was used to make methanol fixed, Giemsa stained slides for each sample at each timepoint. The

trophozoite stage slides were subsequently used to normalise the relative light units (RLU) to 1×10^6 parasites.

Luciferase detection for the FL and all the deletion constructs $\Delta 1-3$ emulated the expected temporal transcription profile peaking at mature trophozoite stage (Fig. 7-12). However, *Luc* activity was significantly reduced in $\Delta 2$ and $\Delta 3$ (Fig. 7-12). This trend was particularly evident at the mature trophozoite stage (temporal peak) where $\Delta 2$ and $\Delta 3$ had reductions in peak luciferase expression of 35% and 55% respectively ($P > 0.01$). For $\Delta 3$, there was also a significant reduction in *luc* activity at the schizont stage of 55% ($P > 0.01$) (ANOVA, with Tukey's post-test, Graphpad Prism 5.0, USA). These data suggest that although the timing of transcript expression is unaffected by the deletion of ORF adjacent sequence, absolute levels of reporter gene are affected when in excess of 200bps of ORF adjacent 5' upstream sequence is deleted.

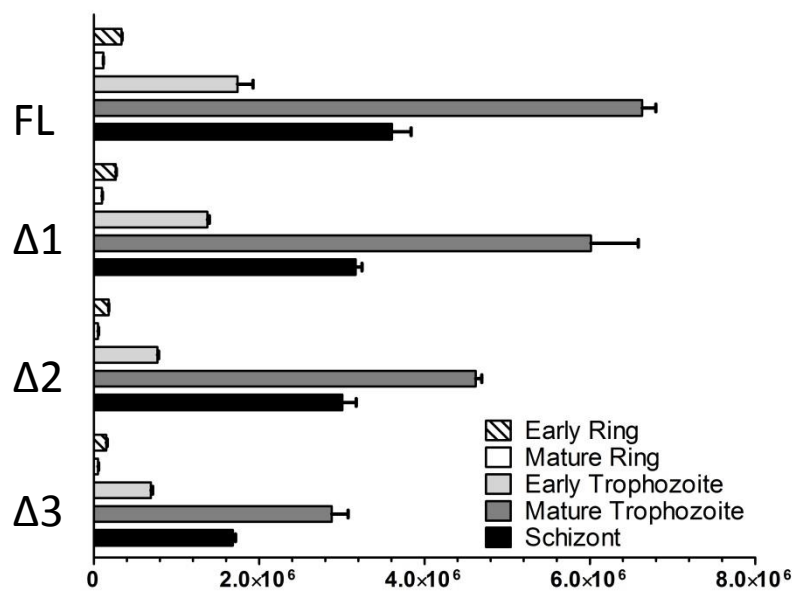


Figure 7-12 Stage-specific luciferase assay

The graph plots the mean \pm stdev (n=9 relative light expression (RLU) per 1×10^6 parasites for each transfectant at five time-points during intraerythrocytic development.

7.7 DISCUSSION

The main aim of this investigation was to understand the effect that the deletion of various lengths of proximal 5' UTR had upon the temporal and absolute levels of expression for a typical *P. falciparum* IE housekeeping gene. Using a *luc* reporter construct and full and matched PFD0660w 5' and 3' upstream and downstream flanking sequences demonstrated that you can delete at least half of the proximal 5' flanking region (c.3/4 of the 5' UTR) and this has no discernible effect on the temporal expression profile of this gene during intraerythrocytic development. Therefore, it seems likely that the components responsible for the timing of expression must reside either: upstream of this region - the area around the putative promoter region was unaffected in $\Delta 1$, $\Delta 2$ or $\Delta 3$ and the region between 661 and 1287bp contains ~50% of all the predicted AP2 binding sites (PlasmoDB) and probably most other promoter related *cis*-acting sites (see Appendix E-5). Or, perhaps more likely, within the 3' UTR which remained unaltered in all transfectants ($\Delta 1$ -3). It is also plausible that the timing of expression is not orchestrated solely by upstream or downstream *cis*-acting sequences. The constitutive assembly of the PIC within intergenic regions (Gopalakrishnan *et al.*, 2009) and the global activity of RNA Pol II (Sims *et al.*, 2009) may also play their part in the continuous cascade of gene expression exhibited by *P. falciparum*. This observation is, however, comparable to previous investigations of the core promoter region of other genes which similarly show no changes in temporal expression (Horrocks *et al.*, 2009), (Wong *et al.*, 2011).

The effect of deleting proximal flanking sequence on absolute levels of transcription and translation looks more subtle. In order to quantify transcription, quantitative RT-PCR was carried out on the four transfected lines (FL and $\Delta 1$ -3) at two time-points; early to mid-trophozoite and mature trophozoite/schizont. This work was carried out in the laboratory after completion of my experimental work. Referring to these data (shown in

Appendix E-4), in conjunction with the luciferase assay output; if you delete 225bps of 5' ORF adjacent sequence ($\Delta 1$) this appears not to affect either transcription or translation. When you delete 416bps of 5' ORF adjacent sequence ($\Delta 2$) you see no effect on transcription but a significant decrease in translation during maximal expression life-cycle stages. However, when you delete 661bps of 5' ORF adjacent sequence ($\Delta 3$) you see a decrease in transcription (a significant decrease of 46.5% at mature trophozoite/schizont stage $p < 0.05$) and a significant decrease in translation during maximal expression life-cycle stages. It should be noted that all comparisons were relative to the FL construct. These data suggest that the reduction in the size of the 5' ORF adjacent intergenic sequence affects predominantly translation (when 416bps upwards are removed) but also appears to effect transcription when over 670bps are removed.

Previous studies deleting 5' UTR regions have shown far more profound effects on the level of expression than demonstrated here (Porter, 2002; Polson and Blackman, 2005; Horrocks *et al.*, 2009). This disparity may be attributable to the approach adopted here. Specifically; i) the use of stably maintained parasite lines which should assemble chromatin and therefore be subject to the appropriate genetic and epigenetic regulatory mechanisms and ii) the use of full and matched 5' and 3' flanking regions enabling all apparent elements necessary for the control of gene expression to be accurately reconstituted (the reporter gene demonstrated the correct temporal profile and all transfectants ($\Delta 1-3$) used the same transcription start and stop sites). One issue I was unable to resolve was an accurate mapping of the transcription start site (TSS) with 5' RACE. The northern blot data suggested that this was in the region of 825bp upstream of the ORF. This issue was also addressed after I had completed my laboratory work by the creation a new construct $\Delta 4$ (Appendix E-3). $\Delta 4$ was identical to all other constructs except it had 943bp of 5' proximal sequence deleted. The $\Delta 4$ transfectant was shown to maintain a stable episomal plasmid population, but produced no *luc* transcript or luciferase expression suggesting the absence of the TSS and indicating that the TSS does

indeed reside between 943 and 661bp upstream of the ORF. Note this transfectant could only be produced in the absence of the pINT plasmid, and therefore contained only stably-maintained episomal plasmid.

That it took the removal of 670bp of 5' proximal sequence to significantly affect the reporter gene expression level strongly suggests an element of length plasticity in the long *P. falciparum* UTR. It is possible that a reduction in the level of expression is only observed when the UTR becomes so short (~200bp) that it interferes with the assembly and processivity of the large RNA polymerase II complex and all its co-factors. It is also possible that such a situation may be 'masked' by the presence of the reporter gene on an episomal plasmid rather than being, more ideally, integrated into the genome. What was also somewhat surprising was that the deletion of the proximal 225bp of sequence - which contained the homopolymeric dA and dT tracts of interest, had no apparent effect upon transcription or translation at all. This leaves the function of these proximal poly dA.dT tracts - in terms of the control of gene expression, unresolved. It is possible, though it seems unlikely, that as the whole intergenic region is punctuated with such poly dA.dT tracts that, as some are removed other poly dA.dT enriched sequences located further upstream replace them (AT content within intergenic regions in *P. falciparum* can be as high as 90%). Or, perhaps, as presented within the previous chapter, the positional coincidence of these ORF adjacent homopolymeric poly dA.dT tracts with nucleosome free regions (NFRs) could suggest that the stably maintained episomal plasmid population may not be subject to identical nucleosomal 'packaging' mechanisms as genomic DNA.

The results from the luciferase assay - trending towards reduced protein levels at maximal life-cycle expression stages with increased loss of 5' proximal flanking sequence suggests that a reduction in length of the 5' proximal UTR may influence the efficiency of translation. Whether this is attributable to the presence of specific 'motifs' or tracts within this sequence

or is simply the by-product of a reduced 5' UTR length it is unfortunately not possible to discern from these data. Although control of protein translation in *P. falciparum* has been documented, particularly mRNA repression during gametogenesis (Mair *et al.*, 2006; Mair *et al.*, 2010) and is evident from the time-lag observed between transcription and translation (Le Roch *et al.*, 2004; Shock *et al.*, 2007) little is known about the actual translational process in this parasite. Eukaryotic translation is generally thought to rely upon the methyl-7-guanosine cap (which is present in *P. falciparum*) and the presence of secondary structure formed by GC rich sequence - identified by the 40S subunit complex. Such GC rich sequences are rare in this AT-rich genome and do not appear to have positional bias - at least relative to the ORF (unpublished observation). The presence of an upstream open reading frame is one plausible explanation - the deletion of which interferes with re-initiation as demonstrated in *var2CSA* (Amulic *et al.*, 2009). Perhaps more likely, however, is that reloading of the ribosomes with elongation factors as they process around the circularised mRNA is less efficient in shorter 5' UTR thus resulting in the production of less protein.

The experimental approach adopted during this study with stably maintained reporter constructs and matched 5' and 3' upstream and downstream sequences is a robust model to provide insights into control of the absolute level of gene expression. Very little detail is still known about the mechanisms of transcriptional and translational control in this parasite and any further data that can be contributed will always be welcome. The inability to alter the temporal expression profile via genetic ablation was somewhat frustrating. Future work would definitely have to entail a systematic evaluation of the 3' downstream sequence. This could be achieved in a very similar manner to the work carried out above, using matched 5' and 3' upstream and downstream sequences respectively, ideally with genomic integration, and the creation of a 3' deletion series to enable evaluation of both temporal and absolute levels of gene expression. These data together, both the 5' and the 3' for the same gene, could

perhaps give a clearer picture of the contribution made by each intergenic region to these processes.

CHAPTER 8 CONCLUSIONS

8.1 INTERGENIC REGIONS AND THE TRANSCRIPTIONAL LANDSCAPE OF *P. FALCIPARUM*

Eukaryotic nuclear genomes vary in size by some 300,000-fold, whereas that of the transcriptome varies only some 17-fold (Cavalier-Smith, 2005). Although it has been demonstrated that a strong correlation exists between the genome and the cell size for vertebrates (Mirsky and Ris, 1951), plants and unicellular protists (Cavalier-Smith, 1985a) a complete lack of correlation is evident between the genome size and the organismal gene quantity or genome complexity (Cavalier-Smith, 1985b). This led the description of the “C-value paradox”, which attempted to address why genome size does not correlate directly with the number of coding sequences (Thomas, 1971). Resolution of this enigma followed in the late 1970s when it was postulated, and subsequently verified, that non-genic DNA (IGR and introns) has function and is therefore subject to specific selective forces that shape genome size and organisation (Cavalier-Smith, 1978; Cavalier-Smith, 1980; Cavalier-Smith, 2005). Interestingly, genome size also correlates with nuclear volume in both animals and plants (Vialli, 1957; Baetcke *et al.*, 1967). Intracellular parasites, in general, have small cell and nuclear volumes and most of the Apicomplexan parasites investigated here have moderately compact genomes of (2.2-4.8kb/ORF). It is therefore likely that their genomes have been subject to selective pressures to improve spatial economy whilst improving metabolic efficiency and coping with rapid cell division which has resulted in the minimization of their genomes through a combination of both gene-loss and reduction in size of non-genic regions (reviewed in Gregory, 2001; Cavalier-Smith, 2005).

In chapter 3, a consensus IGR spacing rule is demonstrated for these moderately compact genomes with the relative size of the three classes of IGR correlating with the nature of the

transcriptional activity that occurs over them. Importantly, whilst the 3:2:1 spacing rule is maintained in these moderately compact genomes, the actual size of the IGR differ between the species investigated – with IGR size correlating with the genome density (or, at least up to 4.8kb/ORF). An important question, therefore, would be why the IGR in *P. falciparum* are so large compared to the other Apicomplexans, i.e. why have the selective forces driving IGR minimization not removed additional IGR? This idea is explored in chapters 4 and 5, and starts with the empirical observation that Northern blot data suggest that *P. falciparum* has relatively long transcripts, much larger than the open reading frame transcribed. Comparison of Northern blot and EST data (chapter 4) and modelling (chapter 5) led to the proposal that UTR lengths typically range between 800-1800 bases and that these are preferentially apportioned at an approximate 75:25 ratio for 5' and 3' UTR, respectively. It is of note that this apportionment directly mirrors that of the IGR space available for type A and C IGR (3:1). Together, work presented in chapters 4 and 5 suggests that the transcriptional landscape over IGR is much more extensive than that suggested solely based on EST data. Indeed, transcriptional units likely spatially overlap within IGR, although are much less likely to temporally overlap – resolving potential steric issues between processing RNA polymerase II complexes. Thus, evidence presented here suggests, in the case of *P. falciparum* at least, that the size of the UTR appears to be a contributing factor in determining the extent of IGR minimization. Further exploration of this idea will require (i) longer RNASeq reads from *P. falciparum* (work ongoing in the Wellcome Trust Sanger Institute) to confirm the modelling of UTR size and apportionment described here, but also (ii) development of similar data in other Apicomplexan parasites to correlate IGR and UTR size.

It is of note that the 3:2:1 consensus IGR spacing rule collapsed when exploring the subtelomeric regions in *P. falciparum*. This region contains members of multigene families implicated in evasion of the human immune system and pathology of disease. Whilst no evidence is presented here, it is conceivable that selection pressures driving IGR

minimization are balanced by the demand to facilitate monoallelic expression and recombination amongst these gene families to ensure a chronic infection in the face of the human host immune response. Interestingly, these longer IGR may not adversely impact on the selective pressure to fit a genome into the available nuclear volume as subtelomeric regions are typically modified into more compact heterochromatin (Westenberger *et al.*, 2009; Ponts *et al.*, 2010).

In terms of the size and apportionment of UTR, data presented here would indicate; (i) that UTR are long - typically some 800–1800 bases and (ii) that 70-80% of the UTR is preferentially apportioned 5' of the ORF. This would suggest that transcriptional start and stop sites lie between 600-1350 bp and 200-450 bp either side of the ORF. Apart from increasing current understanding of the extent of the transcriptional landscape in *P. falciparum*, these more distal transcriptional coordinates have implications for our search and validation of regulatory *cis*-acting regions. *In silico* searches for sequence motifs enriched in the flanking regions of functionally related and/or co-transcribed genes typically use 1kbp of flanking sequence (Elemento *et al.*, 2007; Young *et al.*, 2008). Whilst this would seem suitable for searching downstream of an ORF, it is perhaps not sufficient to identify all potential 5' positioned regulatory elements. To explore whether the revised predictions presented here would lead to additional information about *cis*-acting regions, a search using the FIRE algorithm (finding informative regulatory elements), Elemento *et al.*, (2007), was undertaken. Two windows of 5' flanking sequence, the first between 0 – 1000bp upstream of the ORF and the second between 500-1500bp upstream of the ORF were secured. In total, four datasets of sequence (termed Groups A to D) were created for the analyses. Two captured the 0-1000bp window (A and B) and two the 500-1500bp window (C and D). The difference between the data contained in these pairs of groups reflected the outcome of an upstream flanking ORF encounter. Essentially, all sequence captured within all these datasets was truncated if an upstream ORF was encountered. However, groups A and C (termed ALL)

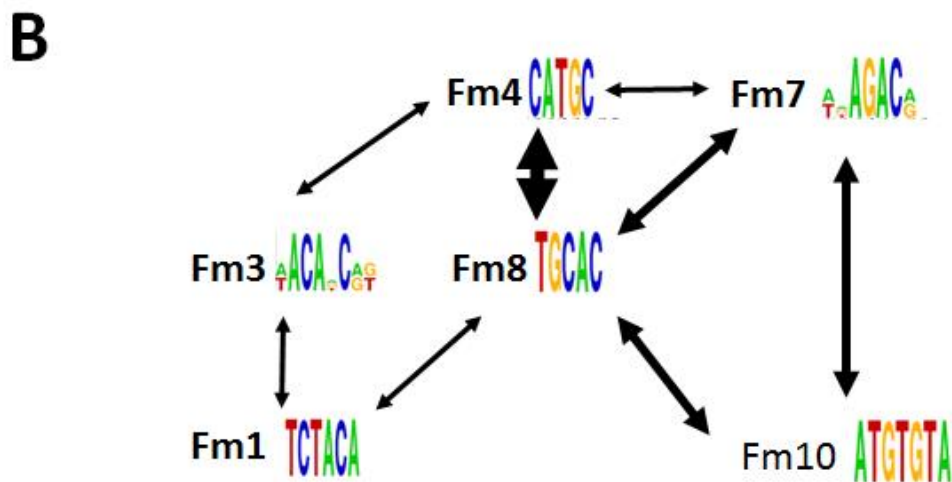
contained all genomic sequences, truncated or not, therefore, all sequences were not of the full 1,000bp, whereas, groups B and D contained only sequences of 1,000bp in length (termed 1000bp). These variations allowed different, but related, groups of sequences to be investigated. The selectivity of the algorithm is then based on the mutuality of *cis*-motifs within the sequences captured with the temporal transcription of the associated gene (Elemento *et al.*, 2007). These analyses captured a total of 27 motifs – with 14 overlapping with the original Elemento study (see Appendix F-8). Being unable to completely recapitulate the Elemento dataset was perhaps not entirely unexpected as different genome builds were used and different sensitivity and stringency settings are available for the algorithm. Disappointingly, only four new motifs (all of which had a low mutual information score) were secured by moving the window of investigation 500bp further upstream – effectively cutting short this avenue for further investigation. What was apparent, however, was the repeated discovery of 11 motifs (Fm1-11, Fig. 8-1A) which occurred in at least two of the groups of sequences investigated (Information taken from Appendix F-1, 3, 5 and 7A). These 11 motifs were all identified in the previous study, had a mutual information (MI) score of 0.041 or above (singletons had a MI score of 0.028 ± 0.004) and 8 of them had a motif readily identifiable as having a cognate Ap2 trans-acting partner in *P. falciparum* (Campbell *et al.*, 2010; Painter *et al.*, 2011) (Fig. 8-2).

Figure 8-1 FIRE motifs common to all datasets and motif interaction map

(overpage) (A) shows the FIRE motifs in common to at least two datasets, their motif label (Fm), their maximum mutual information (MI) score, their corresponding Api-AP2 binding motif and Api-AP2 PlasmoDB gene identifier (Elemento *et al.*, 2007, Campbell *et al.*, 2010, Painter *et al.*, 2011). (B) shows the interaction of co-located motif upstream of co-transcribed genes for Fm1-11 (where the interaction was noted in at least two datasets). This is a composite compiled from the motif interaction heatmaps produced by FIRE (Elemento *et al.*, 2007, Appendices F-2, 4, 6 and 7B). The weight of the connecting line indicates the strength of the relationship as judged based on observation in either: 2, 3 or all 4 groups of sequences investigated.

A

	Max. MI score	Group A (0-1000/all)	Group B (0-1000/1000)	Group C (500-1500/all)	Group D (500-1500/1000)	AP2 binding motif	PlasmoDB
Fm1	0.050					TCTA _c AA _a	Pf3D7_1239200
Fm2	0.105					ATATA _a TA _c	Pf3D7_1222400
Fm3	0.041						
Fm4	0.042					TGCATGC _a	Pf3D7_1466400
Fm5	0.046					C _a CACAC _c C _a CACAC _c T _c CACAC _c	Pf3D7_1456000
Fm6	0.054						Pf3D7_0802100
Fm7	0.121						
Fm8	0.091					GTCAC _T A	Pf3D7_1007700
Fm9	0.042					T _c C _c ACC _a	Pf3D7_1408200
Fm10	0.133						
Fm11	0.041					TAGAACAA	Pf3D7_1139300



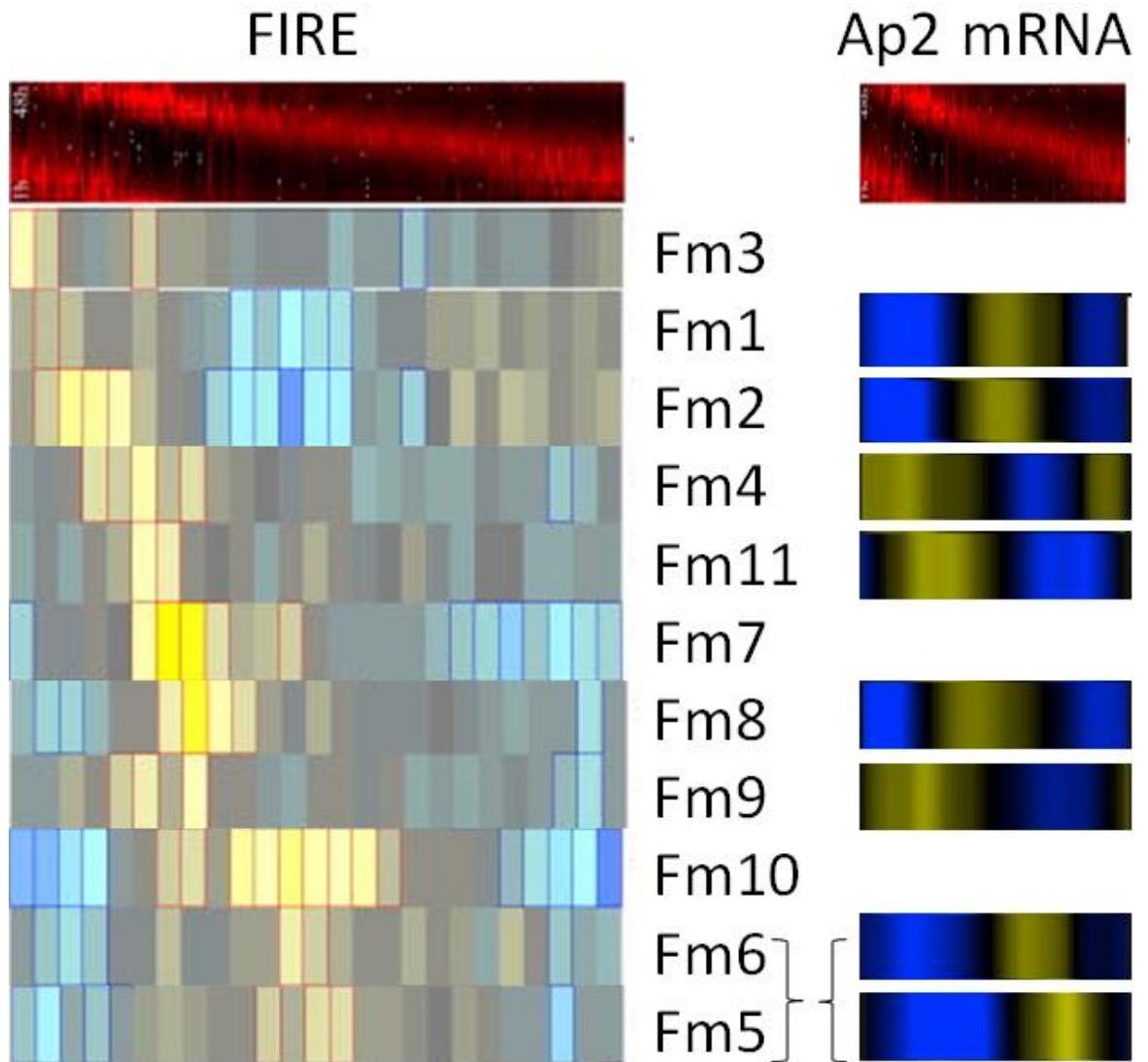


Figure 8-2 Motif heatmap of expression throughout the intraerythrocytic cycle with Api-AP2 transcript profiles.

FIRE motif heatmap (left). High temporal expression is denoted in yellow whereas low temporal expression is indicated in blue (Elemento *et al.*, 2007). The 48 hour *P. falciparum* life-cycle runs left to right. **Motif labels** are shown in the centre (see Fig. 8-1 right for motif). **Api-AP2 mRNA accumulation profiles** are shown on the left where they could be mapped to a motif. High temporal expression is denoted by yellow whereas low temporal expression is denoted by blue. The 48 hour *P. falciparum* life-cycle runs left to right (Campbell *et al.*, 2010, Painter *et al.*, 2011). Note the mRNA transcription/motif correlation for Fm 4, 8, 9 and 11 – all linked into the interaction network in Figure 1B.

Taking motifs Fm1-11, two intriguing features were evident, with implications for our understanding of *cis-trans* promoter based control. First, the FIRE analysis provides heat maps that correlate the mutual association of each motif in the 5' flanking region of genes that share the same temporal pattern of transcription (information taken from Appendix F2, 4, 6 and 7B). Determining the relative association of these motifs with one another, and how often they occurred in each of the four groups of sequences investigated (excluding observations found in only one of groups A-D), provided evidence for an apparent co-ordinating network of *cis*-acting motifs in temporally co-transcribed genes (Fig. 8-1B). This network of *cis*-acting motifs is reminiscent of the combinatorial control of transcription model hypothesized by van Noort and Huynen (2006) which suggested that the utilization of combinations of *cis-trans* motifs could provide the necessary complexity to drive a cascade of temporal gene expression during intraerythrocytic development with only a limited number of specific transcription factors, i.e. the 27 Ap2 transcription factors discovered to date (van Noort and Huynen, 2006). Secondly, closer analysis of the over-representation of these putative *cis*-acting motifs also identified that they appear predominantly to be associated with temporal transcription in the first half of the intraerythrocytic cycle (Fig. 8-2). This observation fuels the debate regarding the interplay of molecular mechanisms that provide for temporal control of transcription. As mentioned previously, prior to 2006, a *cis-trans* promoter based model was favoured, however, as further evidence came to light, more global mechanisms associated with RNA polymerase activity and mRNA half-life were favoured - particularly as they appeared to support the observed trophozoite-stage induction of transcription of a large cohort of the genome (Bozdech *et al.*, 2003; Le Roch *et al.*, 2003; Shock *et al.*, 2007; Sims *et al.*, 2009). This hypothesis, however, still leaves the temporal control of gene expression earlier in the erythrocytic cycle unaccounted for. The evidence presented by Elemento *et al.*, (2007) and the analyses presented here are suggestive of a combinatorial network of putative *cis-trans* interactions associated primarily with early-transcribed genes.

Together, these data suggest that different molecular mechanisms for driving stage-specific transcription may be at play at different life-cycle stages of intraerythrocytic development. 1) Ring stage expression is perhaps more likely based on promoter-based *cis-trans* interactions utilizing Api-AP2 specific transcription factors. 2) Whereas, during trophozoite stage expression other factors are likely at play, such as the global up-regulation of RNAPol II activity and the lengthening of mRNA half-life (Shock *et al.*, 2007; Sims *et al.*, 2009). In support of this hypothesis Wong *et al.*, (2011), using the trophozoite expressed *Pfpcna* gene, created a truncated deletion series. Whilst reporter gene expression identified changes in the absolute levels of reporter gene activity no effect on the temporal expression profile was observed (Wong *et al.*, 2011). In addition, changes in absolute levels of reporter gene activity were observed upon the truncation of the 5' UTR of the trophozoite expressed PFD0660w, but again, no effect was observed upon the temporal expression profile (Chapter 7). In order to understand whether two different primary transcriptional control methodologies are indeed being employed at different time-points during the IE cycle would require promoter deletion studies of ring stage genes to be carried out (preferably using matched 5' and 3' flanking sequence and genomic integration of a single copy of the reporter gene) with the list of high scoring ring-stage FIRE motifs providing a useful candidate list to prioritise genes for analysis.

8.2 WHAT IS THE FUNCTION OF POLY dA.dT TRACTS IN *PLASMODIUM*

SPP.?

The data presented here on the abundance of, and spatial correlation of poly dA.dT tracts with NFRs provides evidence for the role of poly dA.dT tracts as either intrinsic regulators of gene expression through the creation of NFRs, hence enabling access, by extrinsic factors, to the underlying DNA at core promoters, or, as determinants of core nucleosome positioning. In this model, subsequent to genome replication, an ancestral nucleosome would be deposited

after the poly dA.dT tract 'signal' and nucleosomal stacking would occur from this point. The latter hypothesis is perhaps more in line with the data presented here as *P. falciparum* TSSs are located much further upstream than the clear spatial arrangement of poly dA.dT tracts surrounding the ORF. However, homopolymeric dA and dT tracts are also present within CDS in *Plasmodium spp.* which seems to be a unique phenomenon to the haemosporidia family. For *P. falciparum*, it is possible that this is simply a facet of the AT-richness of the genome, or perhaps related to the presence of low-complexity regions in many of its proteins. However, other *Plasmodium spp.*, with lower genomic AT-content, also conform to this arrangement and this has been discussed in Chapter six.

A novel, and at this point unsupported, alternative role for poly dA.dT tracts in *Plasmodium spp.* can be extrapolated from recent findings that suggest that homopolymeric tracts, such as these, act as/or are located within origins of replication in budding and fission yeast (Dai *et al.*, 2005; Field *et al.*, 2008). This has particular relevance if we consider DNA replication in the microgamete during gamete maturation in the midgut of the mosquito vector. During this process the haploid microgametocyte undergoes exflagellation which is completed in the astonishingly short time of approximately 10mins - the product being eight flagellated microgametes (Janse *et al.*, 1986; Raabe *et al.*, 2009). Thus, the process of microgamete formation requires an octoploid increase in DNA in 10mins. Assuming a rate of 2kb/min for eukaryotic DNA polymerase processivity, (average rate of replication fork movement) (Stillman, 1996), this would mean that the microgametocyte would require an exceptionally high genomic density of replication origins to establish this feat within this time-frame. Janse *et al.*, estimated that, from the rate of DNA synthesis, in the region of 1300 replication origins would be required for the 18Mbp *P. berghei* genome which would suggest that the larger 22.85Mbp *P. falciparum* genome would likely require even more (Janse *et al.*, 1986; Hall *et al.*, 2005). Therefore, the extensive repository of homopolymeric tracts throughout the genome

of *Plasmodium spp.*, (i.e. not just restricted to the IGR) could actually be selected for through a novel function - such as origins for replication.

8.3 PLASTICITY OF THE UTR

The functional analysis of putative 5' UTR deletions reported in chapter seven represents an example of best practice in the use of reporter genes to study promoter function in *P. falciparum*. Analysis of the 5' regions were carried out using a matched 3' flanking sequence along with an attempt to ensure reporter constructs were placed within the most appropriate chromatin environment using *bxbl*-integrase mediated integration. Unfortunately, we were unable to obtain a complete series of integrated constructs, although this is akin to previous experience in our laboratory with the *Pfpcna* series and in accordance with colleagues in other laboratories (personal communications). This made a complete quantitative analysis of the effect of 5' UTR deletion on translation difficult - as outlined in the discussion in chapter 7. However, the data clearly demonstrate that: i) deletions of reasonably long lengths of 5'UTR had only minimal effect on absolute and no effect on temporal patterns of transcription, ii) increased deletions of 5' UTR eventually affected absolute, but not temporal, transcript levels, and iii) further increased deletions of 5' UTR also appeared to have some effect on the translation of the reporter gene (as determined from luciferase activity). A trophozoite expressed gene was used for these analyses and therefore, the observation that most of the 5' UTR could be deleted without any effect on the timing of expression, as in other studies (Horrocks *et al.*, 2009; Wong *et al.*, 2011), once again supports the hypothesis of a more global transcription event at this life-cycle stage, or suggests that the components necessary for temporal transcription reside within the 3' UTR. However, the fact that such a large proportion of the potentially long 5' UTR can be deleted, with only a relatively small impact upon processes such as transcription and translation, also suggests that there is likely a degree of inherent plasticity in the length of this region.

Whilst this study has gone some way towards describing the size and apportionment of UTR, why the 5' UTR regions of the transcript would be so large and, as shown here, potentially without major function in directing transcription and translation, has raised more questions than it has provided answers. This is perhaps more intriguing based on the supposition we have made above - that the length of the UTR may be impacting on the extent of IGR minimization possible in *P. falciparum*. That is - if several hundreds of bases of 5' UTR can be removed without affecting the level and timing of transcript accumulation and has a negligible effect on translational efficiency - why haven't these been selected against to allow the IGR size to be further reduced? More detailed analyses of the effects of 5' UTR deletions on transcription and translation are clearly required to support our early premise regarding the plasticity of UTR size. However, these studies are challenging to carry out - the AT-richness of the genome and IGR in particular precludes specific PCR amplification of regions to study. Perhaps, exonuclease digestion or Phusion-based recombineering of the regions of interest would be one route to overcome this. Unfortunately, these are still relatively low-throughput methodologies i.e. they have to be carried out on a gene by gene basis. Another opportunity may be made available through increasing access to high throughput sequencing of genomes from a large number of *P. falciparum* isolates. Previous unpublished work by Horrocks demonstrated length polymorphisms when amplifying 5' flanking regions of *Pfpcna* and *gpb130*. Thus, with increasing availability of IGR from geographically-diverse *P. falciparum* genomes, this would afford the opportunity to identify conserved regions within 5' UTR that may provide 'targets' for more detailed investigation.

Finally, and as is a general feature of work on gene regulatory sequences in our field, this study did not address the function of the 3'UTR. Given the critical role of transcript half-life and that transcripts are 'looped' during translation for efficient unloading/loading of ribosomes, functional studies of this 3'UTR region are long overdue. It is noteworthy that no change in temporal expression has been observed in any promoter deletion studies to date -

despite the cascade of IE gene expression demonstrated so elegantly by Bozdech *et al.* in 2003. Perhaps then, the key is to look for these changes in the 3' UTR.

In summary, new RNASeq datasets will provide further information on the transcriptional landscape of *P. falciparum* outside the ORF in the very near future - these data will be invaluable in elucidating the molecular mechanisms that drive temporal patterns of transcription and translation. I would suggest that the key to exploring the role of *cis-trans* promoter-based control of temporal gene expression lies in the careful selection of ring stage gene sets containing high scoring putative *cis*-acting motifs (as suggested above), the use of matched 5' and 3' UTR - ideally with single-copy full genomic integration and finally, and most importantly, a 5' and 3' truncation series should be evaluated. The output from such a study as this has the potential to provide very interesting results.

APPENDICES

APPENDIX A - HORROCKS *ET AL.*, 2009

APPENDIX D – SUPPLEMENTARY FILES CHAPTER 4

D-1 REFERENCE LIST FOR NORTHERN BLOT DATA TAKEN FROM THE LITERATURE

- Alano *et al.* (1996) Structure and polymorphism of the upstream region of the *pfg27/25* gene, transcriptionally regulated in gametocytogenesis of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 79, 207-217.
- Balu *et al.* (2009) Identification of the transcription initiation site reveals a novel transcript structure for *Plasmodium falciparum* *maebl*. *Exp Parasitol*, 121, 110-114.
- Barik *et al.* (1997) Identification, cloning, and mutational analysis of the casein kinase 1 cDNA of the malaria parasite, *Plasmodium falciparum*: stage-specific expression of the gene. *J Biol Chem*, 272, 26132-8.
- Berry C, *et al.* (1999) A distinct member of the aspartic proteinase gene family from the human malaria parasite *Plasmodium falciparum*. *FEBS Lett*, 447, 149-54.
- Bracchi-Ricard *et al.* (2000) PfPK6, a novel cyclin-dependent kinase/mitogen-activated protein kinase-related protein kinase from *Plasmodium falciparum*. *Biochem J*, 347, 255-63.
- Cheesman *et al.* (1998) Intraerythrocytic expression of topoisomerase II from *Plasmodium falciparum* is developmentally regulated. *Mol Biochem Parasitol*, 92, 39-46.
- Delves *et al.* (1990) Expression of alpha and beta tubulin genes during the asexual and sexual blood stages of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 43, 271-8.
- Dobson *et al.* (2001) Characterization of a novel serine/threonine protein phosphatase (PfPPJ) from the malaria parasite, *Plasmodium falciparum*. *Mol Biochem Parasitol*, 115, 29-39.
- Dobson *et al.* (2003) Characterization of a unique aspartate-rich protein of the SET/TAF-family in the human malaria parasite, *Plasmodium falciparum*, which inhibits protein phosphatase 2A. *Mol Biochem Parasitol*, 126, 239-50.
- Doerig *et al.* (1995) Pfcrk-1, a developmentally regulated cdc2-related protein kinase of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 70, 167-74.

- Doerig *et al.* (1996) A MAP kinase homologue from the human malaria parasite, *Plasmodium falciparum*. *Gene*, 177, 1-6.
- Fan *et al.* (2004a) *Plasmodium falciparum* Histone Acetyltransferase, a Yeast GCN5 Homologue Involved in Chromatin Remodeling. *Euk Cell*, 3, 264-276.
- Fan *et al.* (2004b) PfADA2, a *Plasmodium falciparum* homologue of the transcriptional coactivator ADA2 and its *in vivo* association with the histone acetyltransferase PfGCN5. *Gene*, 336, 251-261.
- Foote *et al.* (1989) Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of *P. falciparum*. *Cell*, 57, 921-930.
- Fox & Bzik (1991) The primary structure of *Plasmodium falciparum* DNA polymerase delta is similar to drug sensitive delta-like viral DNA polymerases. *Mol Biochem Parasitol*, 49, 289-96.
- Hansen *et al.* (2002) A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites. *J Biol Chem*, 277, 4874-82.
- Hicks *et al.* (1991) Glycolytic pathway of the human malaria parasite *Plasmodium falciparum*: primary sequence analysis of the gene encoding 3-phosphoglycerate kinase and chromosomal mapping studies. *Gene*, 100, 123-9.
- Holloway *et al.* (1994) Isolation and characterization of a chaperonin-60 gene of the human malaria parasite *Plasmodium falciparum*. *Mol Biochem Parasitol*, 64, 25-32.
- Horrocks *et al.* (1996) Stage specific expression of proliferating cell nuclear antigen and DNA polymerase delta from *Plasmodium falciparum*. *Mol Biochem Parasitol*, 79, 177-182.
- Horrocks & Kilbey (1996) Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. *Mol Biochem Parasitol*, 82, 207-215.
- Horrocks & Newbold (2000) Intraerythrocytic polyubiquitin expression in *Plasmodium falciparum* is subjected to developmental and heat-shock control. *Mol Biochem Parasitol*, 105, 115-125.
- Horrocks *et al.* (2002) Stage-specific promoter activity from stably maintained episomes in *Plasmodium falciparum*. *Int J Parasitol*, 32, 1203-1206.

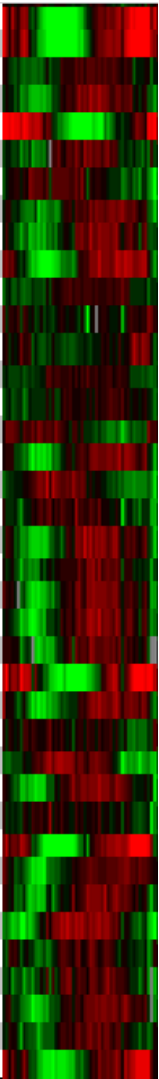
- Joshi *et al.* (1999) Molecular cloning and nuclear localization of a histone deacetylase homologue in *Plasmodium falciparum*. *Mol Biochem Parasitol*, 99, 11-19.
- Knapp *et al.* (1990) *Plasmodium falciparum* aldolase: gene structure and localization. *Mol Biochem Parasitol*, 40, 1-12.
- Krnajski *et al.* (2002) Thioredoxin reductase is essential for the survival of *Plasmodium falciparum* erythrocytic stages. *J Biol Chem*, 277, 25970-5.
- Kyes *et al.* (2002) Stage-specific merozoite surface protein 2 antisense transcripts in *Plasmodium falciparum*. *Mol Biochem Parasitol*, 123, 79-83.
- Kyes *et al.* (2000) A simple RNA analysis method shows *var* and *rif* multigene family expression patterns in *Plasmodium falciparum*. *Mol Biochem Parasitol*, 105, 311-315.
- Lanzer *et al.* (1992a) Transcription mapping of a 100 kb locus of *Plasmodium falciparum* identifies an intergenic region in which transcription terminates and reinitiates. *EMBO J*, 11, 1949-55.
- Lanzer *et al.* (1992b) A sequence element associated with the *Plasmodium falciparum* *kahrp* gene is the site of developmentally regulated protein-DNA interactions. *Nuc Acids Res*, 20, 3051-3056.
- Lanzer *et al.* (1994) Transcriptional and nucleosomal characterization of a subtelomeric gene-cluster flanking a site of chromosomal rearrangements in *Plasmodium falciparum*. *Nuc Acids Res*, 22, 4176-4182.
- Li *et al.* (2004) Isolation and functional characterization of a dynamin-like gene from *Plasmodium falciparum*. *Biochem Biophys Res Commun*, 320, 664-71.
- Li *et al.* (2002) Identification of a second proliferating cell nuclear antigen in the human malarial pathogen *Plasmodium falciparum*. *Int J Parasitol*, 32, 1683-1692.
- Li *et al.* (1989) An enlarged largest subunit of *Plasmodium falciparum* RNA polymerase II defines conserved and variable RNA polymerase domains. *Nuc Acids Res*, 17, 9621-36.
- Li *et al.* (1991) Characterization of the gene encoding the largest subunit of *Plasmodium falciparum* RNA polymerase III. *Mol Biochem Parasitol*, 46, 229-39.

- Martin *et al.* (2000) Characterization of *Plasmodium falciparum* CDP-diacylglycerol synthase, a proteolytically cleaved enzyme. *Mol Biochem Parasitol*, 110, 93-105.
- Mello *et al.* (2002) A multigene family that interacts with the amino terminus of *Plasmodium* MSP-1 identified using the yeast two-hybrid system. *Euk Cell*, 1, 915-925.
- Myrick *et al.* (2003) Mapping of the *Plasmodium falciparum* multidrug resistance gene 5' upstream region, and evidence of induction of transcript levels by antimalarial drugs in chloroquine sensitive parasites. *Mol Microbiol*, 49, 671-683.
- Olafsson *et al.* (1992) Molecular analysis of *Plasmodium falciparum* hexokinase. *Mol Biochem Parasitol*, 56, 89-101.
- Osta *et al.* (2002) A 24bp *cis*-acting element essential for the transcriptional activity of *Plasmodium falciparum* CDP-diacylglycerol synthase gene promoter. *Mol Biochem Parasitol*, 121, 87-98.
- Patterson S, *et al.* (2002) Molecular characterization and expression of an alternate proliferating cell nuclear antigen homologue, PfPCNA2, in *Plasmodium falciparum*. *Biochem. Biophys. Res. Comm.*, 298, 371-376.
- Prasartkaew *et al.* (1996) Molecular cloning of a *Plasmodium falciparum* gene interrupted by 15 introns encoding a functional primase 53 kDa subunit as demonstrated by expression in a baculovirus system. *Nuc Acids Res*, 24, 3934-41.
- Przyborski *et al.* (2003) The histone *H4* gene of *Plasmodium falciparum* is developmentally transcribed in asexual parasites. *Parasitol Res*, 90, 387-9.
- Spielmann & Beck (2000) Analysis of stage-specific transcription in *Plasmodium falciparum* reveals a set of genes exclusively transcribed in ring stage parasites. *Mol Biochem Parasitol*, 111, 453-8.
- Syin & Goldman (1996) Cloning of a *Plasmodium falciparum* gene related to the human 60-kDa heat shock protein. *Mol Biochem Parasitol*, 79, 13-9.
- Tosh *et al.* (1999) *Plasmodium falciparum*: stage-related expression of topoisomerase I. *Exp Parasitol*, 91, 126-32.

- Tosh & Kilbey (1995) The gene encoding topoisomerase I from the human malaria parasite *Plasmodium falciparum*. *Gene*, 163, 151-154.
- Volkman *et al.* (1993) Stage-specific transcripts of the *Plasmodium falciparum* *pfmdr-1* gene. *Mol Biochem Parasitol*, 57, 203-212.
- Waller *et al.* (2003) Chloroquine resistance modulated *in vitro* by expression levels of the *Plasmodium falciparum* chloroquine resistance transporter. *J Biol Chem*, 278, 33593-601.
- Watanabe (1997) Cloning and characterization of heat shock protein DnaJ homologues from *Plasmodium falciparum* and comparison with ring infected erythrocyte surface antigen. *Mol Biochem Parasitol*, 88, 253-8.
- Wellems & Howard (1986) Homologous genes encode two distinct histidine-rich proteins in a cloned isolate of *Plasmodium falciparum*. *Proc Natl Acad Sci USA*, 83, 6065-9.
- Wesseling *et al.* (1989) Stage-specific expression and genomic organization of the actin genes of the malaria parasite *Plasmodium falciparum*. *Mol Biochem Parasitol*, 35, 167-76.
- White *et al.* (1993) The gene encoding DNA polymerase-alpha from *Plasmodium falciparum*. *Nucl Acids Res*, 21, 3643-3646.
- Wickramarachchi *et al.* (2008) Identification and characterization of a novel *Plasmodium falciparum* merozoite apical protein involved in erythrocyte binding and invasion. *Plos One*, 3, e1732.
- Wilson *et al.* (1989) Amplification of a gene related to mammalian *mdr* genes in drug-resistant *Plasmodium falciparum*. *Science*, 244, 1184-1186.
- Zhao *et al.* (1993) Gene structure and expression of an unusual protein-kinase from *Plasmodium falciparum* homologous at its carboxyl terminus with the EF hand calcium-binding proteins. *J Biol Chem*, 268, 4347-4354.

D-2 MICROARRAY DATA PROVIDING PEAK TRANSCRIPT DATA FOR THE 105 GENES
IN THE NORTHERN BLOT COHORT

PlasmoDB ID	Max. Hr.	Life-cycle Stage	IDC Expression Profile
MAL13P1.174	53	S	
MAL13P1.185 (PF13_0206)	24	T	
MAL7P1.27	13	R	
PF07_0064	39	S	
PF07_0065	41	S	
PF08_0034	25	T	
PF08_0131	34	T	
PF10_0016	28	T	
PF10_0084	34	T	
PF10_0143	46	S	
PF10_0153	19	T	
PF10_0154	35	S	
PF10_0160	18	T	
PF10_0165	30	T	
PF10_0193	49	S	
PF10_0362	38	S	
PF11_0039	53	S	
PF11_0061	32	T	
PF11_0117	34	T	
PF11_0271	34	T	
PF11_0282	39	S	
PF11_0338	47	S	
PF11_0377	48	S	
PF11_0425	37	S	
PF11_0457	53	S	
PF11_0465	47	S	
PF11_0486	53	S	
PF13_0011	49	S	
PF13_0150	18	T	
PF13_0193	5	R	
PF13_0199	42	S	
PF13_0200	37	S	
PF13_0291	34	T	
PF13_0328	40	S	
PF14_0053	33	T	
PF14_0078	18	T	
PF14_0097	24	T	
PF14_0124	16	R	
PF14_0148	39	S	
PF14_0149	48	S	
PF14_0177	38	S	
PF14_0224	48	S	
PF14_0254	32	T	
PF14_0295	39	S	
PF14_0316	50	S	
PF14_0323	46	S	
PF14_0425	53	S	
PF14_0602	36	S	
PFA0285c	46	S	
PFA0290w	39	S	
PFA0545c	39	S	
PFB0115w	16	R	
PFB0295w	12	R	

PFB0300c	48	S	
PFB0815w	48	S	
PFB0840w	36	S	
PFB0895c	33	T	
PFC0090w	53	S	
PFC0340w	31	T	
PFC0805w	12	R	
PFD0590c	40	S	
PFD0595w	36	S	
PFD0660w	42	S	
PFD0865c	41	S	
PFD1050w	53	S	
PFD1105w	48	S	
PFE0285c	18	T	
PFE0520c	40	S	
PFE1150w	14	R	
PFE1345c	42	S	
PFE1590w	16	R	
PFF1155w (MAL6P1.189)	16	R	
PFF1225c (MAL6P1.175)	30	T	
PFI0155c	32	T	
PFI0180w	34	T	
PFI0235w	33	T	
PFI0530c	31	T	
PFI0540w	49	S	
PFI0725c	34	T	
PFI0740c	18	T	
PFI1105w	20	T	
PFI1170c	34	T	
PFI1260c	15	R	
PFI1475w	47	S	
PFI1665w	42	S	
PFL1285c	32	T	
PFL1545c	22	T	
PFL1655c	23	T	
PFL1790w	38	S	
PFL1915w	34	T	
PFL2005w	34	T	
PFL2215w	48	S	

Microarray data providing peak transcript data for the 105 genes in the Northern Blot cohort. The unique PlasmoDB gene identifier is indicated with the reported peak in transcript accumulation from the Malaria intraerythrocytic development cycle (IDC) strain comparison database (<http://malaria.ucsf.edu/comparison>) A phaseogram image is shown for each gene. Red represents transcript accumulation and green a relative underrepresentation (Bozdech *et al.*, 2003). A life-cycle stage was annotated for peak transcript accumulation where; Ring (R) is 1-16hrs post infection (hpi), trophozoite (T) 17-32hpi and schizont (S) >35hpi

APPENDIX E – SUPPLEMENTARY FILES CHAPTER 7

APPENDIX E-1 ORF ADJACENT HOMOPOLYMERIC DA AND DT TRACTS

Upstream Sequence

```
>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2) | genomic |  
Pf3D7_04_v3 forward | (geneStart-1286 to geneStart+0) | length=1287  
ACCATACACAAATAAACATAGACACAGATACCTCCACCCACAATACATATATAAATAGCA  
ACAAACAACTAATACTTTCCCTTACCAATTTAACAAGTCTACATATAACAATATCTAGAT  
ACACACACATATCTCTCAATCTAAATAAATGGCATGTAAAATACACCAACATAACTATGA  
TATATTATATTATATAATAAAATTAATGTTATGTTATATATGTAAAGATAAAAAATCT  
TATATATAAAGAGAATAGTAAAATTTAAAATATAAAAAATGTGATTTATTTTTTAAATA  
TGTTTATATTTGAAAGGGGAGAAAAGATATAAAATAAAATGAGAAAATAATTATATTAATA  
AAAATATTATGATTTTTAATAAAATTTTATATTATATACATATAAATATTTGAT  
TGTTATACATTTCTTTTTGATATATATTTTTTATAAGTGTTTGTGTTAAATTAATA  
ATAAATATAAATATAATTTGTGGTTAAAATAAAAAATGATAAATGAAATAATATATATA  
TATATATTTTATAAATAAAATTAAGATAAAAATAAATCTAAGGTATTTAGTATATTTTTG  
GTTTTCACTATAAATATTATAATGATATATATATATATATATATATTTTTTTTTTTTT  
TTTTTTTTTTTTTTTTCATATGTTGAGTAAAAAATATGTATGTGAATACACAGAACCA  
TAAAATAAATTTTTAAGAATAATAAATAAATAATGAATGAATTTTACTTTTTTAAATAT  
ATTTTTGTATTTGTTAAATAAATAATTATTATTGATTATATATATATATATTTTT  
TTTTGGTGAGGATTTAGTGC GTTAAATATTTGCATATTTCAATATAAATTGAAATAA  
TATAATACATATATATATATATATTTTTTTTTTTTTTTTTTATATTTAATATTAAT  
TTTAAATGTTTATTTTATATATTTTAAATTAATTTCTTTTAAATATTTATATTAT  
TTTTAATTTTTTTATGTTTCAGATGTGTATATTTTACACATTTTTTTATTGATTTT  
TTTTTTTTTTTTTTTTTTATGTAATAAATTGTTGTAATTTATTAAATTGGATTT  
CATAAATATACACATATATTACAAAATATATATATATTATATATATATATATATATAT  
ATGATATATTTTTATTTATTTTTTTTTATTTGTTAAAATAAAAAAGTGTGTGAGAATATA  
TATAAAAAAAAAAATATGAGTAAAAA
```

Downstream Sequence

```
>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2) | genomic |  
Pf3D7_04_v3 forward | (geneEnd+0 to geneEnd+674) | length=675  
ATAATTCCTTTGAATTATGA AAAAAAAAAAATATATATATTTACATAGAATATATAAATC  
TATTAATATGATAAATGGTAAATATATATATATATAAATATATATGTATGTATATAT  
GTATATATGTGTATGCTA TTTTTTATTTTTATTTTATATATTTTAAAAACACTA  
GATAATTTATTTAATTATATATCTCTTATATTTTTTTTTAAATATTATTAATAAATACA  
TAATAAATTTTAAACGGAATTATAAAAAAAAAATTTATTTGTTTATATATATATATAT  
ATATATATATATATATATATATTACGTTAAAAGAAAAGAAAAAAGAAAAAAGATAATA  
TTGTTGATAGTTAATATTAGATGGATGTAATTTTTGTATAAATCCATATAAATATTTTAC  
TTTTCAATATAAAAAAAAAAAAAAAAAAAAAAAAAATTTGATTTTAAACAATATAAATATCTATA  
TAATAATAAAAAAAAAAAAAAAAAAGATGCCACAATAAAAAATATCAGTTGAGCAGATTTTA  
TAGTCATATAAATATGTGTGTTTGAACCAATCCTGTGTCATTAAAATAAATATAAACAAT  
ATATATATATATATATACATATATGATATTTTTTATTATTATATTTCTGTAATCAT  
TATAACATTCGTTAT
```

APPENDIX E-2 CLONING PRIMER LOCATIONS

Upstream Sequence

>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2)
| genomic | Pf3D7_04_v3 forward | (geneStart-1286 to geneStart+0) |
length=1287

ACCATACACAAATAAACATAGACACAGATACCTCCACCCACAATACATATATAAAATAGCA P1
ACAAACAAACTAATACTTTCCCTTACCAATTTAACAAGTCTACATATAACAATATCTAGAT
ACACACACATATCTCTCAATCTAAATAAATGGCATGTAAAATACACCAACATAACTATGA
TATATTATATTATATATAATAAAAATTTAAATGTTATGTTATATATGTAAAGATAAAAAATCT
TATATATAAAGAGAATAGTAAAATTTAAAATATAAAAATGTGTATTTATTTTTTTAAATA
TGTTTTATATTTGAAAGGGGAGAAAGATATAAAAATAAAATGAGAAATAATTATATTTAAATA P31
AAAATATTATATGATTTTTTAATAAAAATTTTTATATTATATATACATATAAAATATTTGTAT
TGTTATACATTTCTCTTTTTGATATATATATTTTTTATAAGTGTTTTTGTAAATTTAAAT
ATAAATATAAATATAATTTGTGGTTAAAAATAAAAAATGATAAATATGAAAATAATATATATA
TATATATATTTATAATAAAAATTAAGATAAAAATAAAATCTAAGGTATTTAGTATATTTTTTG P3
GTTTTCACTATAATATTATAATGATATATATATATATATATATATATTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTCATATGTTGAGTAAAAAATATGTATGTGAATACACAGAACCA
TAAAAATAATATTTTAAGAATAATAATAAATAATGAATGAATTTACTTTTTTTAAATAT
ATATTTTGTATTTTTGTTAAATAAATAATTATTATTATTGATTATATATATATATATATTTT
TTTTTTGGTGAGGATTTAGTGCGTTAAATATTTTGCATATTTCAATATAAAATGAAATAA P4
TATAATACATATATATATATATATATTTTTTTTTTTTTTTTTTTTTTATTTTTTTATATTTTAAATATAAT
TTTTAAATGTTTATTATTTATATATATTTTTAAATTTATATTCCTTTAAAAATATTATATTAT
TTTTAATTTTTTTTATGTTTCAGATGTGTTATATTTTTTACACAATTTTTTTTATTGATTTTTT P5
TTTTTTTTTTTTTTTTTTTTTTTATGTAATAAATTGTTGTAATAATTATTATTAATTTGGATTT
CATAAATATACACATATATTACAAAATATATATATATATTATATATATATATATATATATAT
ATGATATATTTTTTATTTATTTTTTTTTTTTTTTTATTTGTTAAAAATAAAAAGTGTGTGAGAATATA P2
TATAAAAAAATAAATATGAGTAAAAA

Gene Sequence

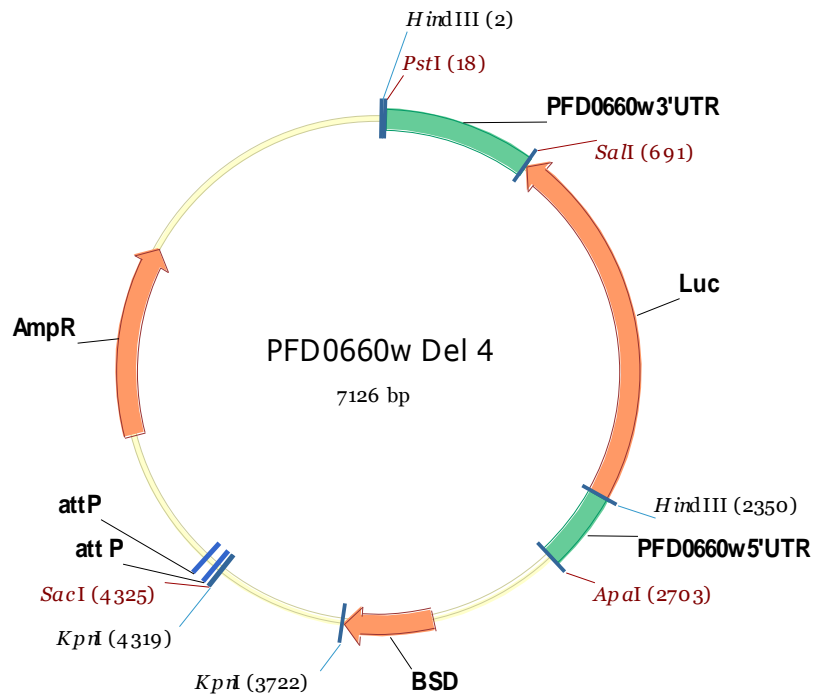
>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2)
| genomic | Pf3D7_04_v3 forward | (geneStart+0 to geneEnd+0) | length=888

ATGTATGTAAATTTTTTTAATAAATTTAAATATCCCCAAATGTTTAAAGAACAAATTTTTAAC
TTTAAAGAAAGATTTAGTGAAACGAAACAAAATATTCAAATAAAAATATATTTAAAGAAAAAT
AAAAAGAAATATGTATTTTATACAGCCGTTACATCATTATCCTTATCAATAGCGTTGACT
TATGCATATGAGAAATTCGTACTTTATAAATGGAATCGTAGATATGATTATTTATTATAAT
CCAAATATAAAATTTTTTTGAAAAGAAAGAAAAAAGAAAGGTAAGAAAGATAAAAAAGAAAT
ACAACGAAACATATTATATTAGTTAGACATGGACAATATGAAAGAAGATATAAAGATGAT
GAGAATTCATAACGTTTAACTAAAGAAGGGTGTAAACAAGCTGATATAACGGGTAAAAAA
TTAAAAGATATTTTAAATAATAAAAAAGGTTAGTGTTATTTATCATTTCAGATATGATAAGA
GCTAAAGAAACGGCTAATATTATAAGTAAATTTTTTCCCTGATGCTAATTTAATAAATGAT
CCAAATTTAAATGAAGGAACCCCTTATTTACCTGACCCTCTTCCAAGACATTCAAAATTT
GATGCTCAAAAAATTAAGAAGATAATAAAAAGAATAAATAAAGCTTATGAACTTATTTTT
TATAAACCTAGTGGTGATGAAGATGAATATCAATTAGTAATATGTCACGGAAATGTAATT
AGATATTTCTTGTGTAGAGCTCTACAAATTCCTTATTTGCATGGTTACGATTTTCAAGT
TATAATTGTGGTATTACATGGTTAGTTTTAGATGATGAGGGTCTGTAGTTTTAAGAGAA
TTTGGTTCGGTTTCTCACCTCCCCTTTGAAAGTGTAACATATTTTTTAA

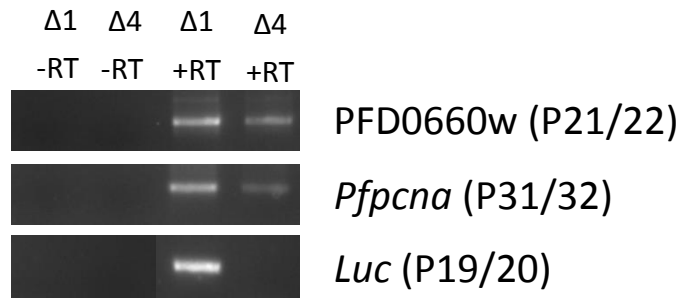
Downstream Sequence

>PF3D7_0413500 | Plasmodium falciparum 3D7 | phosphoglucomutase-2 (PGM2)
| genomic | Pf3D7_04_v3 forward | (geneEnd+0 to geneEnd+674) | length=675
ATAATTCCTTTGAATTATGAAAAAAAAAAAAATATATATATTACATAGAAATATATAAATC P8
TATTAATATGATAATATGGTAAATATATATATATATATAAATATATATGTATGTATATAT
GTATATATGTGTGTATGCTATTTTTTTTTATTTTTTATTTTTATATATTTTTTAAAACTA
GATAATTTATTTTAATTATATATCTCTTATATATTTTTTTTTAAATATTATTAATAAATACA
TAATAAATTATTTAAACGGAATTATAAAAAAAAAATTTATATTGTTTATATATATATATATAT
ATATATATATATATATATATATATATTACGTTAAAAGAAAAGAAAAAAAAAGAAAAAGATAATA
TTGTTGATAGTTAATATTAGATGGATGTAATTTTGTATAAATCCATATAAATATTTAC
TTTTCAATATAAAAAAAAAAAAAAAAAAAAAAAAAATTTGATTTTAACAATATAAATATCTATA
TAATAATAATAAAAAAAAAAAAAAAAAAGATGCCACAATAAAAAATATCAGTTGAGCACATATTA
TAGTCATATAAATATGTGTGTTTGAACCAATCCTGTGTCATTAAAAATAAATATAATACAAT
ATATATATATATATATATATACATATATGTATATTTTTTTATTTATTATATTCTGTAAATCAT P9
TATAACATTCGTTAT

APPENDIX E-3 Δ4 PLASMID FOR TRANSFECTION AND PROOF OF INTEGRATION

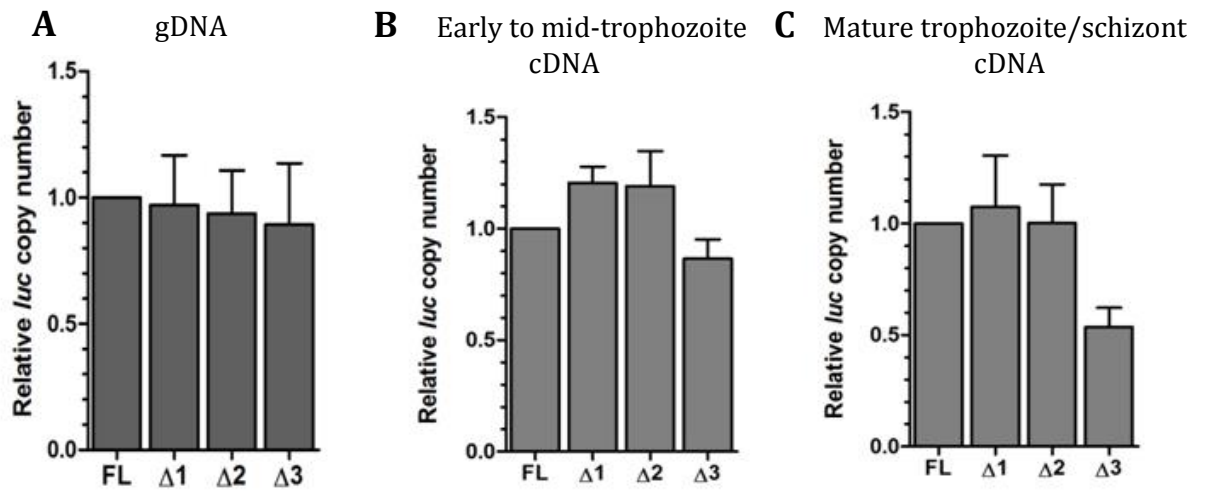


A further deletion construct was created after the completion of my practical work by Dr Sandra Hasenkamp from the Horrocks lab in order to try and resolve the dilemma of the transcription start site. This Δ4 plasmid contained a 343bp 5' UTR (944bp of FL ORF flanking 5' UTR was removed from this construct which included our predicted putative transcription start site). The luciferase reporter gene and the FL 3' UTR were identical to all other constructs.



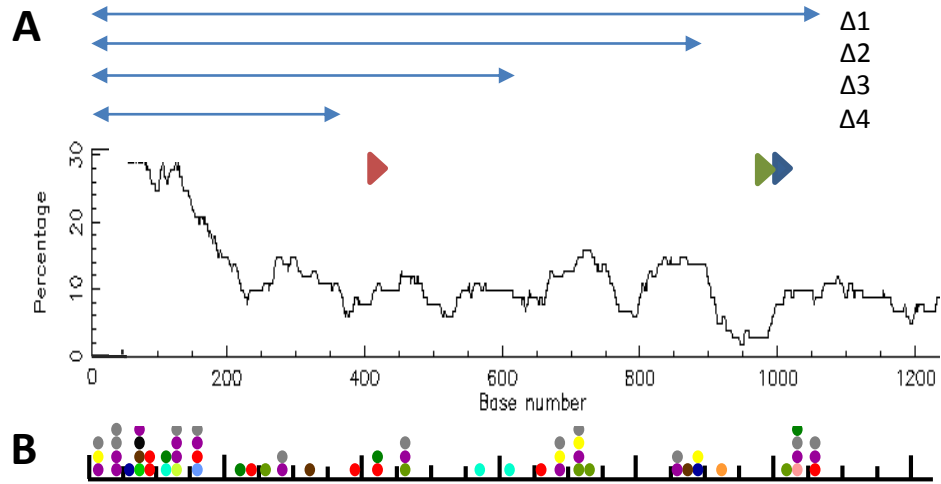
RT-PCR confirmation of absence of *luc* transcription in Δ4 transfectant. Mixed stage total RNA was reverse transcribed using random-hexamer primers (+RT). As a control, reactions omitting reverse transcriptase (-RT) were used. Using the indicated primers, the presence of PFD0660w transcripts and those of a second control gene (*P. falciparum* proliferating cell nuclear antigen PF13_0328) are present in both Δ1 and Δ4 transfectants. Whilst Δ1 contains *luc* transcripts, these cannot be detected in Δ4.

APPENDIX E-4 RT-PCR



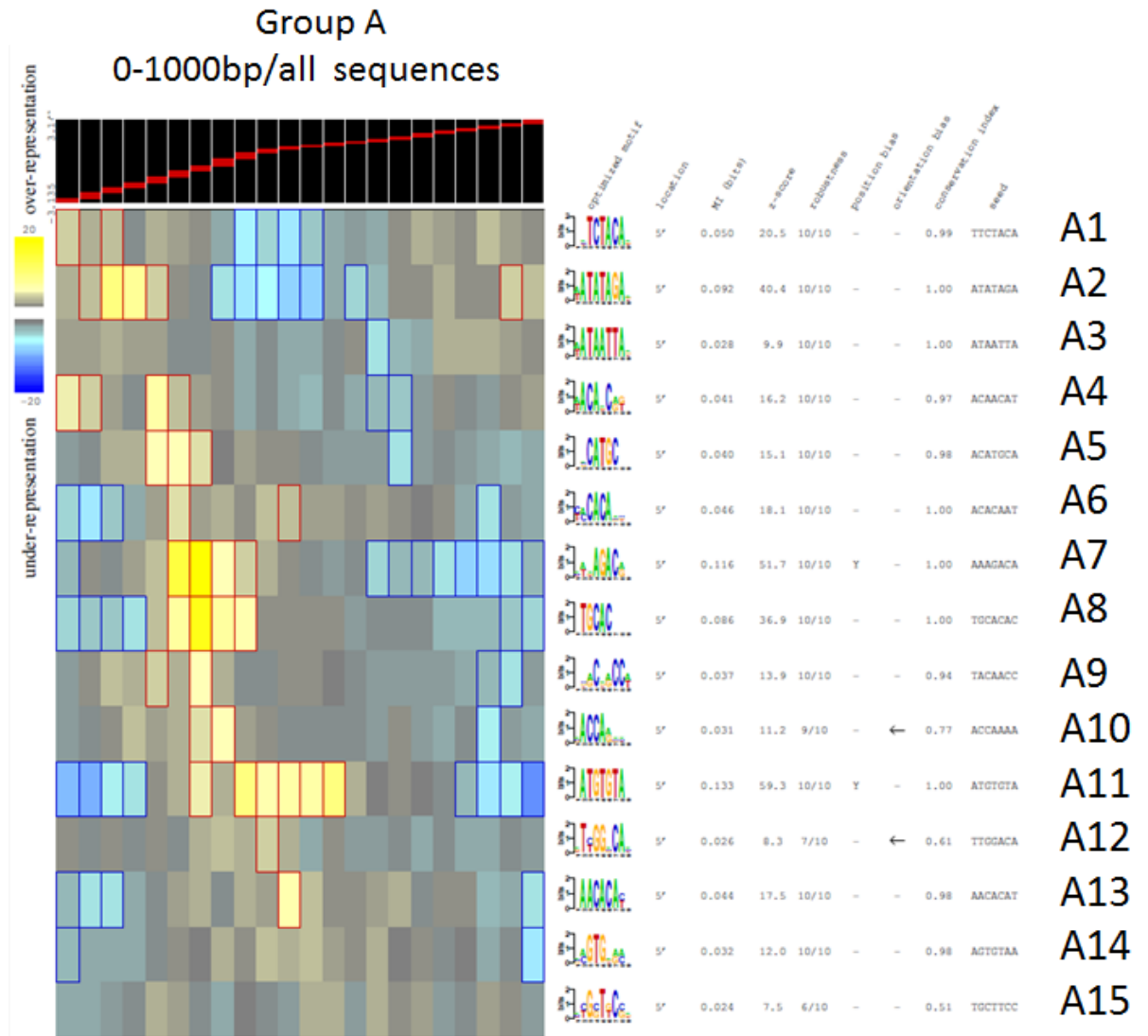
Quantitative PCR analysis of relative *luc* gene copy numbers in indicated *luc* transfectants. The graph depicts a mean normalised *luc* count (to seryl-tRNA synthetase) relative to the FL transfectant \pm StDev ($n = 3$, $p > 0.05$ ANOVA). (B) Quantitative RT-PCR analysis of relative *luc* cDNA copy numbers in indicated *luc* transfectants in early to mid-trophozoites (C) and mature trophozoites. All data are presented relative to the FL transfectant \pm StDev ($n = 4$), * $p > 0.05$ ** $p > 0.01$ on Tukeys multiple comparison test. In (C) $\Delta 3$ is significantly different to all other samples.

APPENDIX E-5 PLOT REPRESENTING %GC OVER THE 5' INTERGENIC REGION OF
PFD0660W AND PREDICTED AP2 BINDING SITES



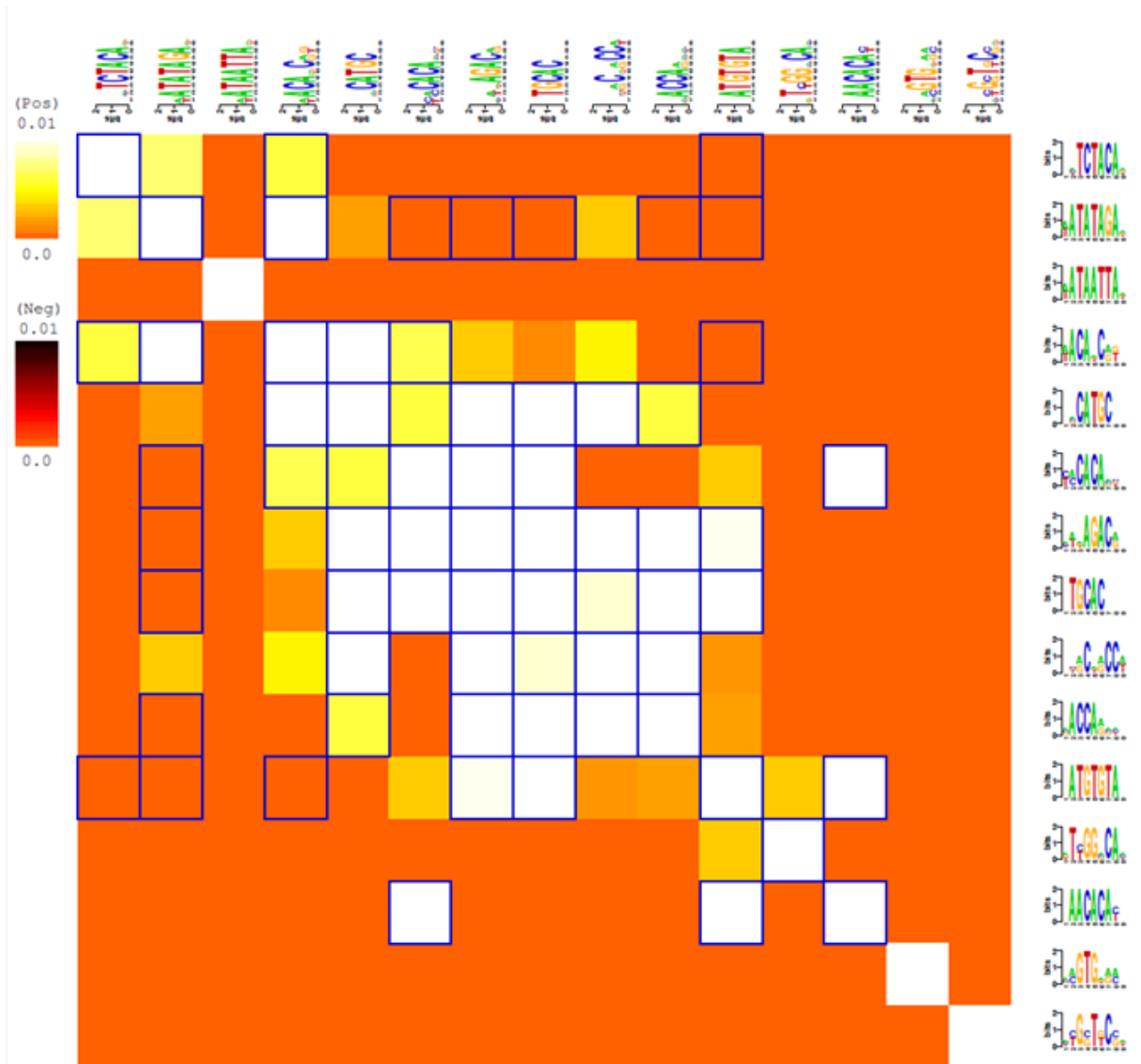
(A) Schematic representing the 1286bp of 5' flanking sequence for PFD0660w. The top part of the panel illustrates the location of the 5' flanking sequence inserted into the Δ series of transfection plasmids. Below is a plot of percentage GC content for this region (<http://www.ebi.ac.uk/Tools/emboss/cpgplot/>). The location of the transcription start sites from existing 5' EST data (blue arrow), the RLM-RACE data reported in this study (green arrow) and that predicted from the length of the UTR from northern blot (red arrow) are indicated on this plot. Note the extreme AT bias in sequences immediately upstream of the RACE sites, where AT content >95%. (B) Schematic illustrating the incidence of predicted AP2-binding sites (from Campbell *et al.* PLoS Pathogens 6, e1001165) in 50bp bins of the 5' flanking sequence. The 16 Ap2 binding sites predicted are each indicated using a different colour dot.

F-1 FIRE ANALYSIS GROUP A

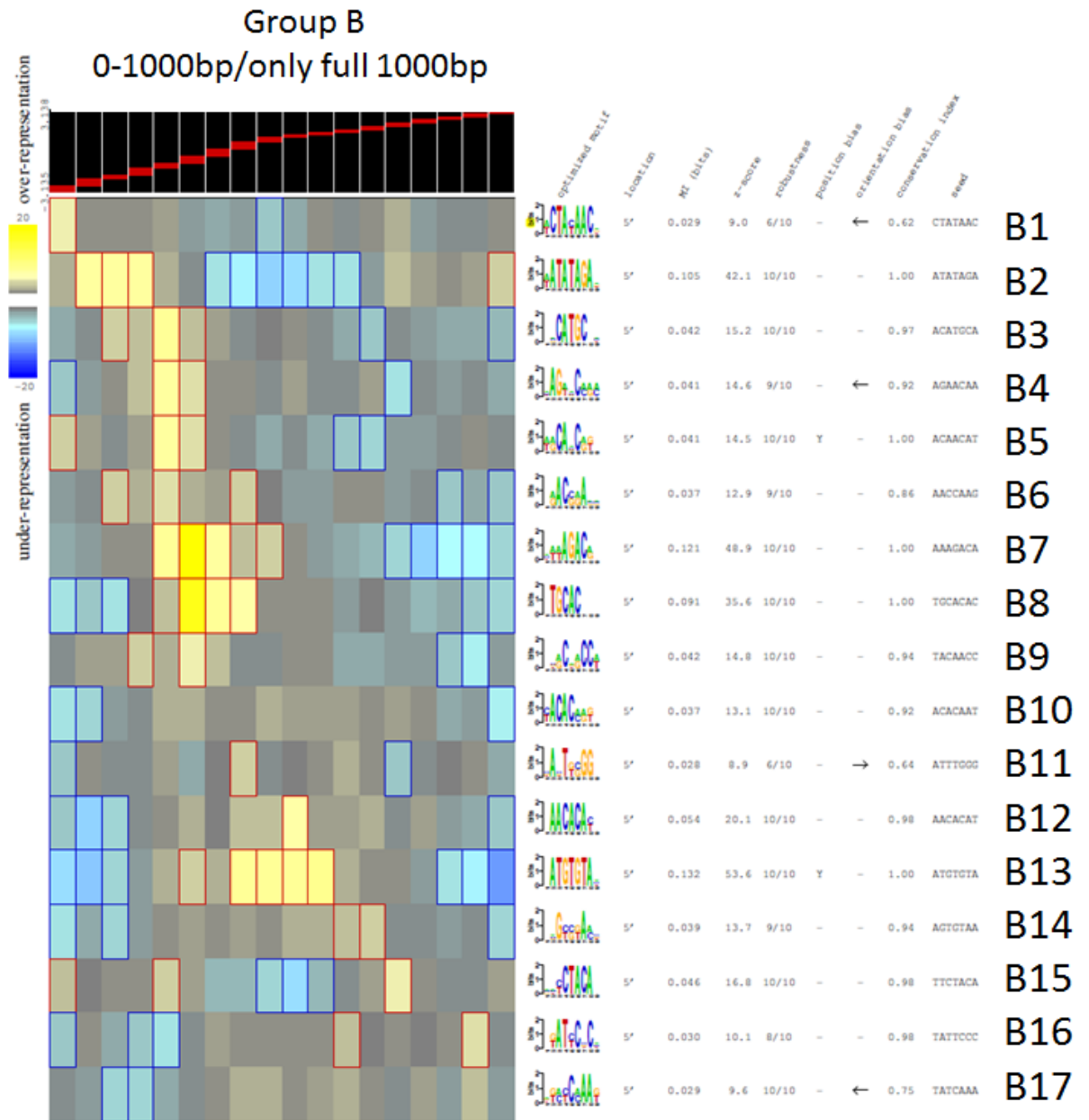


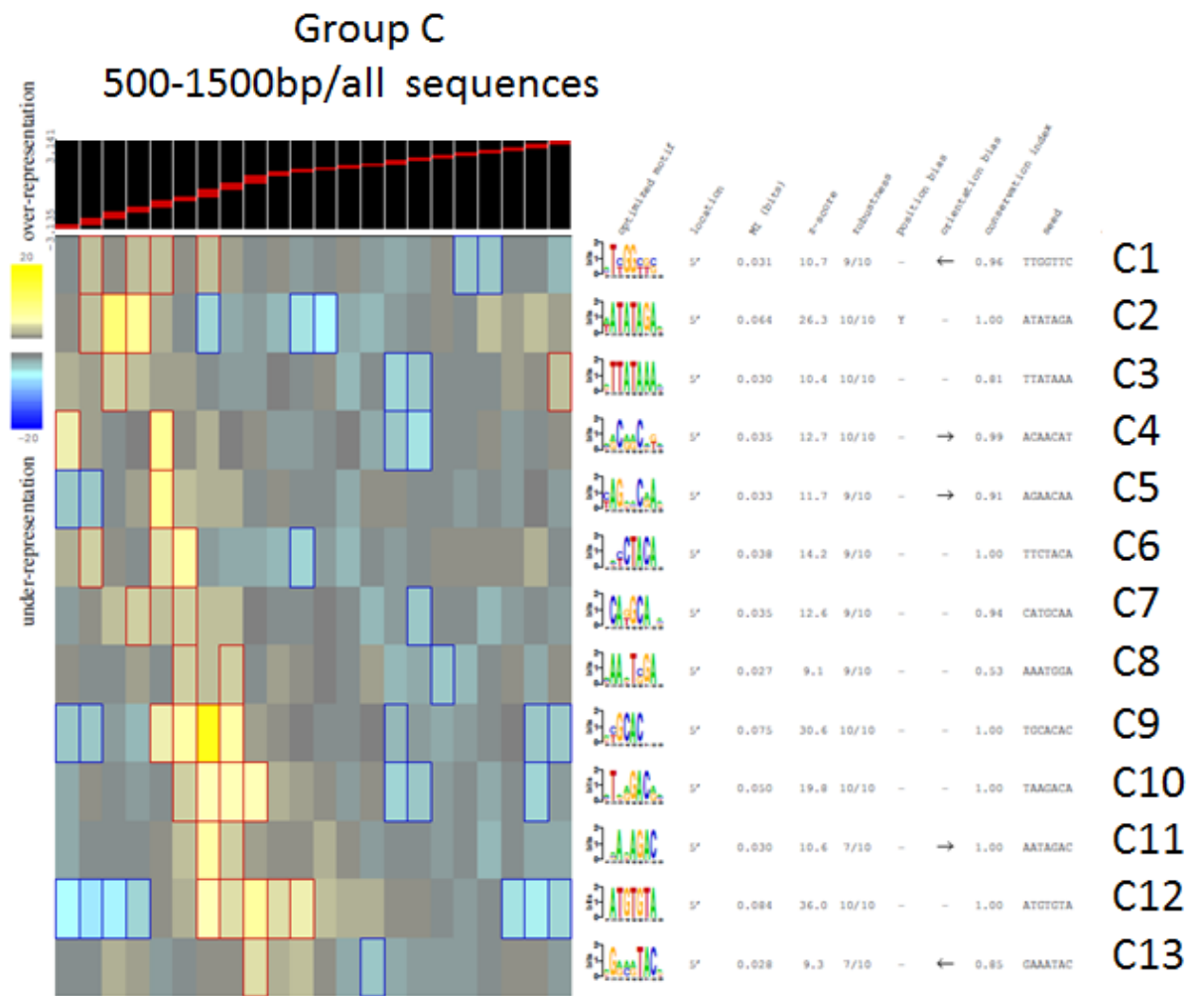
F-2 – INTERACTIONS AMONGST GROUP A MOTIFS

Group A
0-1000bp/all sequences



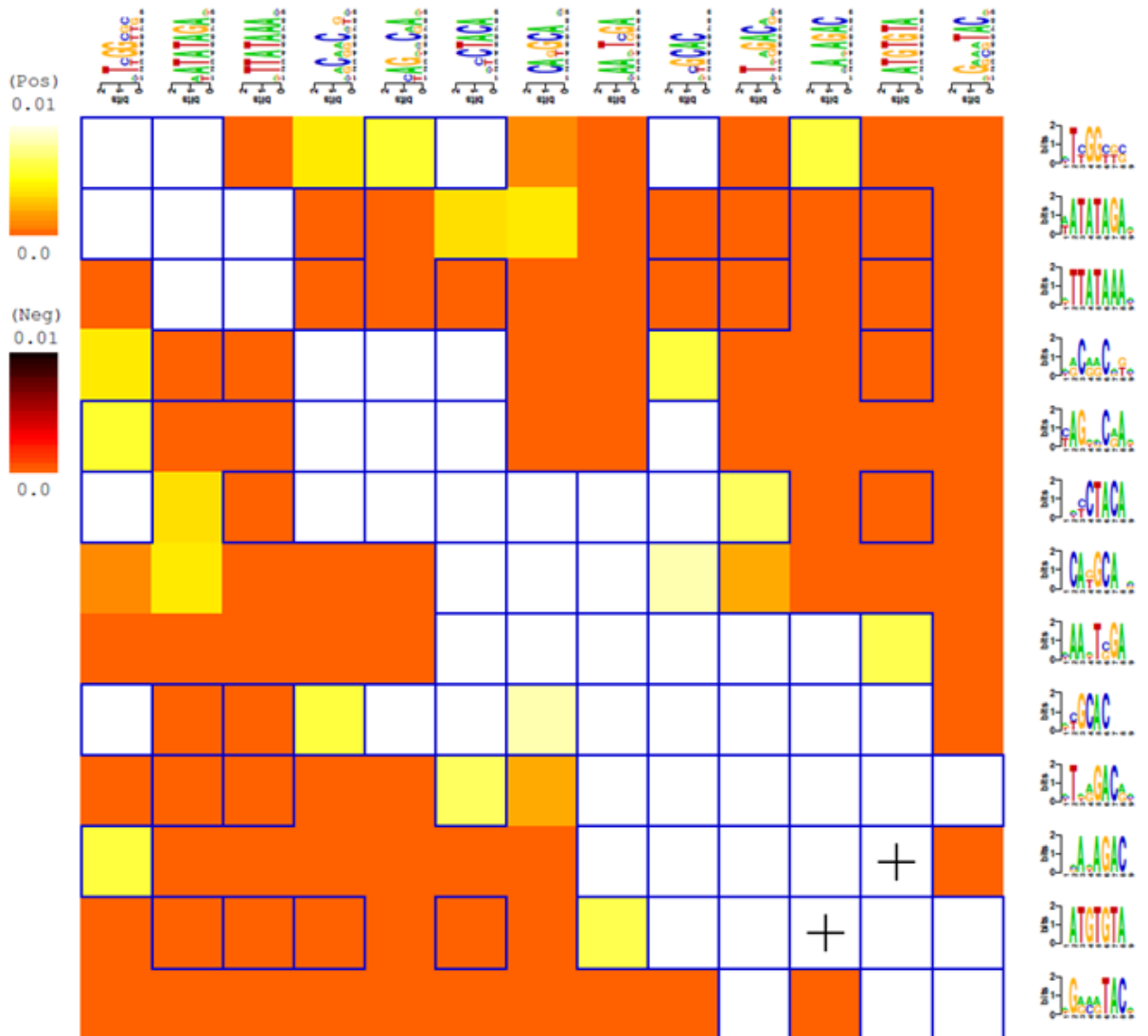
F-3 FIRE ANALYSIS GROUP B



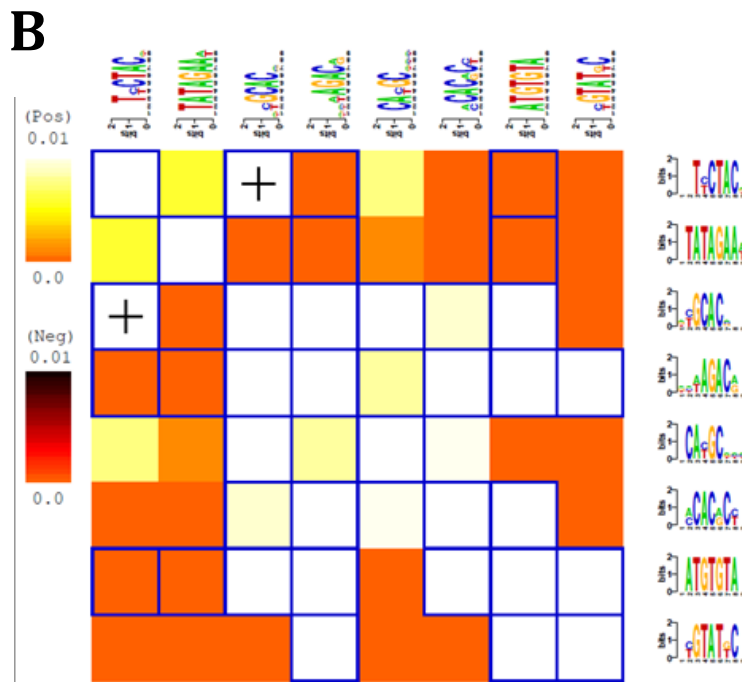
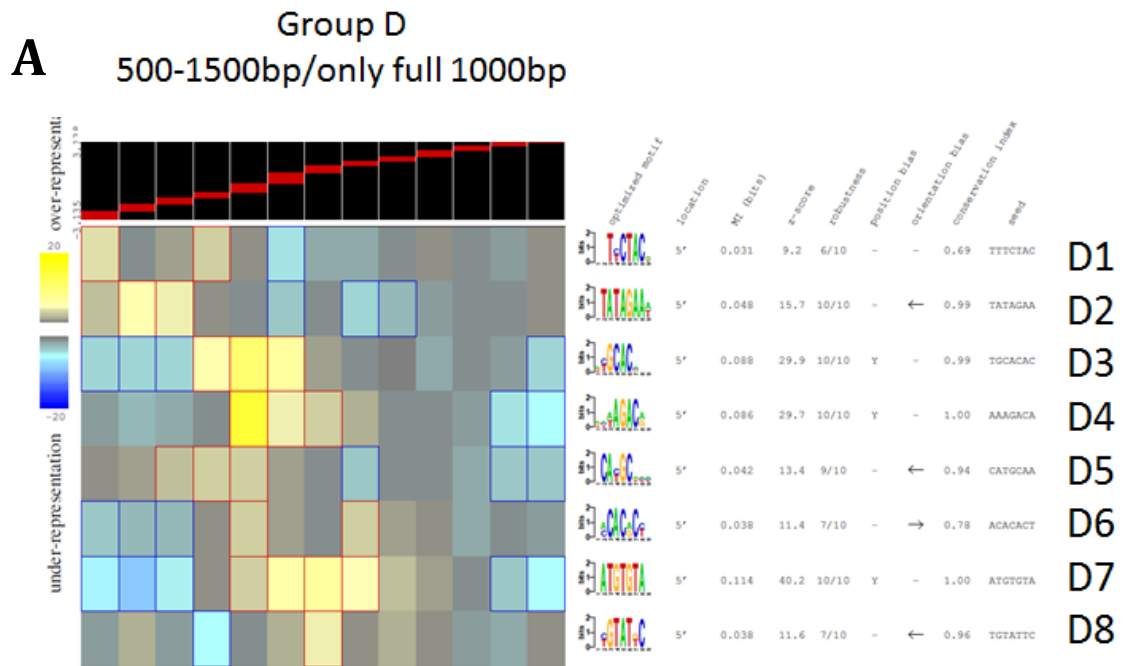


F-6 INTERACTIONS AMONGST GROUP C MOTIFS

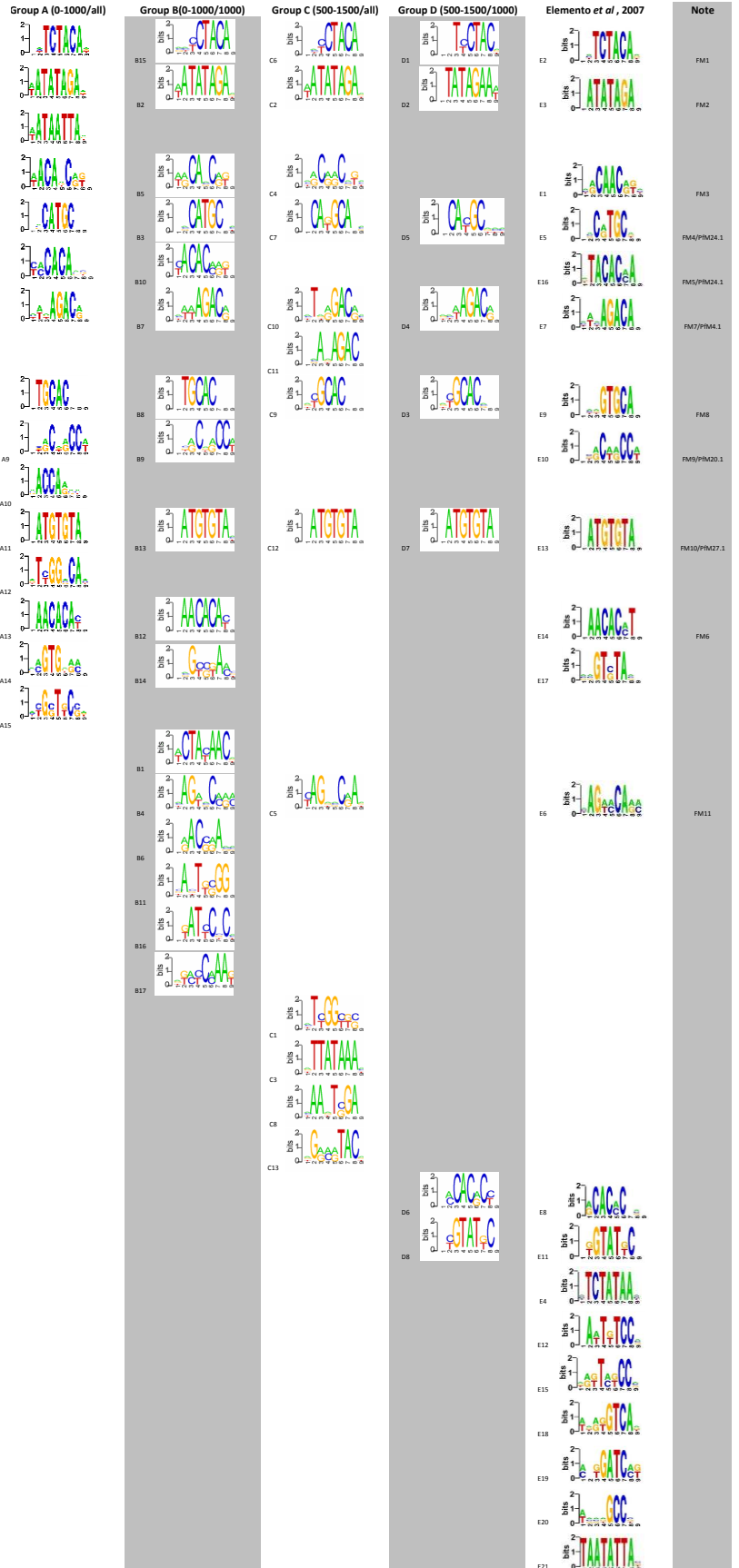
Group C
500-1500bp/all sequences



F-7 FIRE ANALYSIS AND INTERACTION AMONGST MOTIFS GROUP D



F-8 FIRE MOTIF
COMPARISON
BETWEEN ALL
DATASETS



Appendices F1-8. **F-1** 0-1000_ALL FIRE motif heat map, **F-3** 0-1000_1000bp FIRE motif heat map, **F-5** 500-1500_ALL FIRE motif heat map and **F-7A** 500-1500_1000bp FIRE motif heat map. The major findings from these analyses were: i) High scoring motifs (captured >1) are more likely to have a cognate Api-AP2 binding protein identified - suggesting a higher degree of confidence in the prediction (Campbell *et al.*, 2010, Painter *et al.*, 2011,) and ii) importantly, most of the high scoring motifs identified here were associated with intraerythrocytic schizogony early transcribed genes. Six strong motif contenders were identified in all four datasets ATATAGA, ATGTGTA, AGAC, TGCAC, TCTACA and CATGC. ATGTGTA is perhaps the exception as this motif appears to be prominent slightly later during the IE life-cycle than the rest of the motifs and was associated by Elemento *et al.*, with DNA replication. (Elemento *et al.*, 2007). **F-2** 0-1000_ALL FIRE motif interaction heatmap, **F-4** 0-1,000_1000bp FIRE interaction heatmap, **F-6** 500-1500_ALL FIRE motif interaction heatmap and **F-7B** 500-1500_1000bp FIRE motif interaction heatmap. ATATAGA is the earliest most prominent ring stage motif and it does not appear to interact with any other motif. It does, however, look surprisingly like a TATA-box and has an apparent cognate Api-AP2 binding protein (Campbell *et al.*, 2010, Painter *et al.*, 2011). Multiple interactions appear to occur between the remaining motifs (ATGTGTA, AGAC, TGCAC, TCTACA and CATGC) and these are reasonably consistent between all datasets. The fact that many of the motifs appear to interact with one another is interesting and reminiscent of the combinatorial control of transcription model hypothesized by van Noort and Huynen (van Noort and Huynen, 2006). It is also worth re-iterating here that many of the Api-AP2 specific transcription factor proteins have been demonstrated to have multiple binding sites with varying degrees of affinity for different motifs (Campbell *et al.*, 2010). **F-8** Also of note is that I was unable to entirely recapitulate the original dataset (Elemento *et al.*, 2007) which was likely attributable to differences in the stringency and sensitivity thresholds and/or the fact that the genome version I utilized for these analyses was different to that of Elemento *et al.* However, all high scoring motifs reported by Elemento *et al.*, were also identified in the equivalent 0-1000_ALL dataset.

REFERENCES

- ABKARIAN, M., MASSIERA, G., BERRY, L., ROQUES, M. & BRAUN-BRETON, C. 2011. A novel mechanism for egress of malarial parasites from red blood cells. *Blood*, 117, 4118-24.
- AGNANDJI, S. T., LELL, B., FERNANDES, J. F., ABOSSOLO, B. P., METHOGO, B. G., KABWENDE, A. L., ADEGNIKA, A. A., MORDMULLER, B., ISSIFOU, S., KREMSNER, P. G., SACARLAL, J., AIDE, P., LANASPA, M., APONTE, J. J., MACHEVO, S., ACACIO, S., BULO, H., SIGAUQUE, B., MACETE, E., ALONSO, P., ABDULLA, S., SALIM, N., MINJA, R., MPINA, M., AHMED, S., ALI, A. M., MTORO, A. T., HAMAD, A. S., MUTANI, P., TANNER, M., TINTO, H., D'ALESSANDRO, U., SORGHO, H., VALEA, I., BIHOUN, B., GUIRAUD, I., KABORE, B., SOMBIE, O., GUIGUEMDE, R. T., OUEDRAOGO, J. B., HAMEL, M. J., KARIUKI, S., ONEKO, M., ODERO, C., OTIENO, K., AWINO, N., MCMORROW, M., MUTURI-KIOI, V., LASERSON, K. F., SLUTSKER, L., OTIENO, W., OTIENO, L., OTSYULA, N., GONDI, S., OTIENO, A., OWIRA, V., OGUK, E., ODONGO, G., WOODS, J. B., OGUTU, B., NJUGUNA, P., CHILENGI, R., AKOO, P., KERUBO, C., MAINGI, C., LANG, T., OLOTU, A., BEJON, P., MARSH, K., MWAMBINGU, G., OWUSU-AGYEI, S., ASANTE, K. P., OSEI-KWAKYE, K., BOAHEN, O., DOSOO, D., ASANTE, I., ADJEI, G., KWARA, E., CHANDRAMOHAN, D., GREENWOOD, B., LUSINGU, J., GESASE, S., MALABEJA, A., ABDUL, O., MAHENDE, C., LIHELUKA, E., MALLE, L., LEMNGE, M., THEANDER, T. G., DRAKELEY, C., ANSONG, D., AGBENYEGA, T., ADJEI, S., BOATENG, H. O., RETTIG, T., BAWA, J., SYLVERKEN, J., SAMBIAN, D., SARFO, A., AGYEKUM, A., et al. 2012. A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants. *N Engl J Med*, 367, 2284-95.
- AGNANDJI, S. T., LELL, B., SOULANOUDJINGAR, S. S., FERNANDES, J. F., ABOSSOLO, B. P., CONZELMANN, C., METHOGO, B. G., DOUCKA, Y., FLAMEN, A., MORDMULLER, B., ISSIFOU, S., KREMSNER, P. G., SACARLAL, J., AIDE, P., LANASPA, M., APONTE, J. J., NHAMUAVE, A., QUELHAS, D., BASSAT, Q., MANDJATE, S., MACETE, E., ALONSO, P., ABDULLA, S., SALIM, N., JUMA, O., SHOMARI, M., SHUBIS, K., MACHERA, F., HAMAD, A. S., MINJA, R., MTORO, A., SYKES, A., AHMED, S., URASSA, A. M., ALI, A. M., MWANGOKA, G., TANNER, M., TINTO, H., D'ALESSANDRO, U., SORGHO, H., VALEA, I., TAHITA, M. C., KABORE, W., OUEDRAOGO, S., SANDRINE, Y., GUIGUEMDE, R. T., OUEDRAOGO, J. B., HAMEL, M. J., KARIUKI, S., ODERO, C., ONEKO, M., OTIENO, K., AWINO, N., OMOTO, J., WILLIAMSON, J., MUTURI-KIOI, V., LASERSON, K. F., SLUTSKER, L., OTIENO, W., OTIENO, L., NEKOYE, O., GONDI, S., OTIENO, A., OGUTU, B., WASUNA, R., OWIRA, V., JONES, D., ONYANGO, A. A., NJUGUNA, P., CHILENGI, R., AKOO, P., KERUBO, C., GITAKA, J., MAINGI, C., LANG, T., OLOTU, A., TSOFA, B., BEJON, P., PESHU, N., MARSH, K., OWUSU-AGYEI, S., ASANTE, K. P., OSEI-KWAKYE, K., BOAHEN, O., AYAMBA, S., KAYAN, K., OWUSU-OFORI, R., DOSOO, D., ASANTE, I., ADJEI, G., ADJEI, G., CHANDRAMOHAN, D., GREENWOOD, B., LUSINGU, J., GESASE, S., MALABEJA, A., ABDUL, O., KILAVO, H., MAHENDE, C., LIHELUKA, E., et al. 2011. First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *N Engl J Med*, 365, 1863-75.
- ALANO, P. 2007. *Plasmodium falciparum* gametocytes: still many secrets of a hidden life. *Mol Microbiol*, 66, 291-302.
- ALONSO, P. L. & TANNER, M. 2013. Public health challenges and prospects for malaria control and elimination. *Nat Med*, 19, 150-5.
- AMULIC, B., SALANTI, A., LAVSTSEN, T., NIELSEN, M. A. & DEITSCH, K. W. 2009. An upstream open reading frame controls translation of *var2csa*, a gene implicated in placental malaria. *PLoS Pathogens*, 5, e1000256.

- ANSELMINI, C., BOCCHINFUSO, G., DE SANTIS, P., SAVINO, M. & SCIPIONI, A. 1999. Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. *J Mol Biol*, 286, 1293-301.
- ARYA, G., MAITRA, A. & GRIGORYEV, S. A. 2010. A structural perspective on the where, how, why, and what of nucleosome positioning. *J Biomol Struct Dyn*, 27, 803-20.
- BAER, K., KLOTZ, C., KAPPE, S. H. I., SCHNIEDER, T. & FREVERT, U. 2007a. Release of Hepatic *Plasmodium yoelii* Merozoites into the Pulmonary Microvasculature. *PLoS Pathogens*, 3, e171.
- BAER, K., ROOSEVELT, M., CLARKSON, A. B., JR., VAN ROOIJEN, N., SCHNIEDER, T. & FREVERT, U. 2007b. Kupffer cells are obligatory for *Plasmodium yoelii* sporozoite infection of the liver. *Cell Microbiol*, 9, 397-412.
- BAETCKE, K. P., SPARROW, A. H., NAUMAN, C. H. & SCHWEMMER, S. S. 1967. The relationship of DNA content to nuclear and chromosome volumes and to radiosensitivity (LD50). *Proc Natl Acad Sci U S A*, 58, 533-40.
- BALAJI, S., BABU, M. M., IYER, L. M. & ARAVIND, L. 2005. Discovery of the principal specific transcription factors of apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucl. Acids Res.*, 33, 3994-4006.
- BARTFAI, R., CUI, L., HORROCKS, P. & MIAO, J. 2013. Chromatin Structure and Functions (unpublished).
- BARTFAI, R., HOEIJMAKERS, W. A., SALCEDO-AMAYA, A. M., SMITS, A. H., JANSSEN-MEGENS, E., KAN, A., TREECK, M., GILBERGER, T. W., FRANCOIJS, K. J. & STUNNENBERG, H. G. 2010. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLoS Pathog*, 6, e1001223.
- BEIER, J. C. 1998. Malaria parasite development in mosquitoes. *Annu Rev Entomol*, 43, 519-43.
- BEUTLER, E. 1959. The hemolytic effect of primaquine and related compounds: a review. *Blood*, 14, 103-39.
- BIZZARO, J. W. & MARX, K. A. 2003. Poly: a quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. *BMC Bioinformatics*, 4, 22.
- BOPP, S. E., MANARY, M. J., BRIGHT, A. T., JOHNSTON, G. L., DHARIA, N. V., LUNA, F. L., MCCORMACK, S., PLOUFFE, D., MCNAMARA, C. W., WALKER, J. R., FIDOCK, D. A., DENCHI, E. L. & WINZELER, E. A. 2013. Mitotic evolution of *Plasmodium falciparum* shows a stable core genome but recombination in antigen families. *PLoS Genet*, 9, e1003293.
- BOZDECH, Z., LLINAS, M., PULLIAM, B. L., WONG, E. D., ZHU, J. & DERISI, J. L. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biology*, 1, e5.
- BOZDECH, Z., MOK, S., HU, G., IMWONG, M., JAIDEE, A., RUSSELL, B., GINSBURG, H., NOSTEN, F., DAY, N. P., WHITE, N. J., CARLTON, J. M. & PRIESER, P. R. 2008. The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc Natl Acad Sci U S A*, 105, 16290-5.
- BOZDECH, Z. & PRIESER, P. 2013. Functional genomics of *Plasmodium* parasites. In *Malaria parasites: comparative genomics, evolution and molecular biology*. Eds Carlton, J., Perkins, S.L., Deitsch, K.W., Caister Academic Press.
- BRAKS, J. A. M., MAIR, G. R., FRANKE-FAYARD, B., JANSE, C. J. & WATERS, A. P. 2008. A conserved U-rich RNA region implicated in regulation of translation in *Plasmodium* female gametocytes. *Nucleic Acids Research*, 36, 1176-1186.
- BRANCUCCI, N. M., WITMER, K., SCHMID, C. D., FLUECK, C. & VOSS, T. S. 2012. Identification of a cis-acting DNA-protein interaction implicated in singular var gene choice in *Plasmodium falciparum*. *Cell Microbiol*, 14, 1836-48.
- BRICK, K., WATANABE, J. & PIZZI, E. 2008. Core promoters are predicted by their distinct physicochemical properties in the genome of *Plasmodium falciparum*. *Genome Biol*, 9, R178.
- BROADBENT, K. M., PARK, D., WOLF, A. R., VAN TYNE, D., SIMS, J. S., RIBACKE, U., VOLKMAN, S., DURASINGH, M., WIRTH, D., SABETI, P. C. & RINN, J. L. 2011.

- A global transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol*, 12, R56.
- CALLAN-JONES, A., ALBARRAN ARRIAGADA, O. E., MASSIERA, G., LORMAN, V. & ABKARIAN, M. 2012. Red blood cell membrane dynamics during malaria parasite egress. *Biophys J*, 103, 2475-83.
- CAMPBELL, T. L., DE SILVA, E. K., OLSZEWSKI, K. L., ELEMENTO, O. & LLINAS, M. 2010. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathogens*, 6, e1001165.
- CANN, H., BROWN, S. V., OGUARIRI, R. M. & GOLIGHTLY, L. M. 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, 239-45.
- CARLTON, J. M., ADAMS, J. H., SILVA, J. C., BIDWELL, S. L., LORENZI, H., CALER, E., CRABTREE, J., ANGIUOLI, S. V., MERINO, E. F., AMEDEO, P., CHENG, Q., COULSON, R. M., CRABB, B. S., DEL PORTILLO, H. A., ESSIEN, K., FELDBLYUM, T. V., FERNANDEZ-BECERRA, C., GILSON, P. R., GUEYE, A. H., GUO, X., KANG'A, S., KOIJ, T. W., KORSINCZKY, M., MEYER, E. V., NENE, V., PAULSEN, I., WHITE, O., RALPH, S. A., REN, Q., SARGEANT, T. J., SALZBERG, S. L., STOECKERT, C. J., SULLIVAN, S. A., YAMAMOTO, M. M., HOFFMAN, S. L., WORTMAN, J. R., GARDNER, M. J., GALINSKI, M. R., BARNWELL, J. W. & FRASER-LIGGETT, C. M. 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 455, 757-63.
- CARLTON, J. M., ANGIUOLI, S. V., SUH, B. B., KOIJ, T. W., PERTEA, M., SILVA, J. C., ERMOLAEVA, M. D., ALLEN, J. E., SELENGUT, J. D., KOO, H. L., PETERSON, J. D., POP, M., KOSACK, D. S., SHUMWAY, M. F., BIDWELL, S. L., SHALLOM, S. J., VAN AKEN, S. E., RIEDMULLER, S. B., FELDBLYUM, T. V., CHO, J. K., QUACKENBUSH, J., SEDEGAH, M., SHOAI, A., CUMMINGS, L. M., FLORENS, L., YATES, J. R., RAINE, J. D., SINDEN, R. E., HARRIS, M. A., CUNNINGHAM, D. A., PREISER, P. R., BERGMAN, L. W., VAIDYA, A. B., VAN LIN, L. H., JANSE, C. J., WATERS, A. P., SMITH, H. O., WHITE, O. R., SALZBERG, S. L., VENTER, J. C., FRASER, C. M., HOFFMAN, S. L., GARDNER, M. J. & CARUCCI, D. J. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, 419, 512-9.
- CAVALIER-SMITH, T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J Cell Sci*, 34, 247-78.
- CAVALIER-SMITH, T. 1980. r- and K-tactics in the evolution of protist developmental systems: cell and genome size, phenotype diversifying selection, and cell cycle patterns. *Biosystems*, 12, 43-59.
- CAVALIER-SMITH, T. 1985a. Cell volume and the evolution of genome size. In *The evolution of genome size*. Chichester: Wiley 105-184. Ed Cavalier-Smith, T.
- CAVALIER-SMITH, T. 1985b. Eukaryote gene numbers, non-coding DNA and genome size. In *The evolution of genome size*. Chichester: Wiley 69-103. Ed Cavalier-Smith, T.
- CAVALIER-SMITH, T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot*, 95, 147-75.
- CERUTTI, H. & CASAS-MOLLANO, J. A. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Current Genetics*, 50, 81-99.
- CHAAL, B. K., GUPTA, A. P., WASTUWIDYANINGTYAS, B. D., LUAH, Y. H. & BOZDECH, Z. 2010. Histone deacetylases play a major role in the transcriptional regulation of the *Plasmodium falciparum* life cycle. *PLoS Pathog*, 6, e1000737.
- CHAKRABARTI, K., PEARSON, M., GRATE, L., STERNE-WEILER, T., DEANS, J., DONOHUE, J. P. & ARES, M. 2007. Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *Rna-a Publication of the Rna Society*, 13, 1923-1939.
- CHANG, G. S., NOEGEL, A. A., MAVRICH, T. N., MULLER, R., TOMSHO, L., WARD, E., FELDER, M., JIANG, C., EICHINGER, L., GLOCKNER, G., SCHUSTER, S. C. &

- PUGH, B. F. 2012. Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in *Dictyostelium*. *Genome Res*, 22, 1098-106.
- CHOI, S. W., KEYES, M. K. & HORROCKS, P. 2006. LC/ESI-MS demonstrates the absence of 5-methyl-2'-deoxycytosine in *Plasmodium falciparum* genomic DNA. *Mol Biochem Parasitol*, 150, 350-2.
- CHOOKAJORN, T., DZIKOWSKI, R., FRANK, M., LI, F., JIWANI, A. Z., HARTL, D. L. & DEITSCH, K. W. 2007. Epigenetic memory at malaria virulence genes. *Proc Natl Acad Sci U S A*, 104, 899-902.
- COHANIM, A. B. & HARAN, T. E. 2009. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res*, 37, 6466-76.
- COULSON, R. M. R., HALL, N. & OUZOUNIS, C. A. 2004. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Research*, 14, 1548-1554.
- COWMAN, A. F., BERRY, D. & BAUM, J. 2012. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *J Cell Biol*, 198, 961-71.
- CUI, L., FAN, Q., CUI, L. & MIAO, J. 2008. Histone lysine methyltransferases and demethylases in *Plasmodium falciparum*. *Int J Parasitol*, 38, 1083-97.
- CUI, L. & MIAO, J. 2010. Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. *Eukaryot Cell*, 9, 1138-49.
- CURTIDOR, H., VANEGAS, M., ALBA, M. P. & PATARROYO, M. E. 2011. Functional, immunological and three-dimensional analysis of chemically synthesised sporozoite peptides as components of a fully-effective antimalarial vaccine. *Curr Med Chem*, 18, 4470-502.
- DAI, J., CHUANG, R. Y. & KELLY, T. J. 2005. DNA replication origins in the *Schizosaccharomyces pombe* genome. *Proc Natl Acad Sci U S A*, 102, 337-42.
- DAILY, J. P., LE ROCH, K. G., SARR, O., NDIAYE, D., LUKENS, A., ZHOU, Y., NDIR, O., MBOUP, S., SULTAN, A., WINZELER, E. A. & WIRTH, D. F. 2005. In vivo transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. *J Infect Dis*, 191, 1196-203.
- DAILY, J. P., SCANFELD, D., POCHE, N., LE ROCH, K., PLOUFFE, D., KAMAL, M., SARR, O., MBOUP, S., NDIR, O., WYPIJ, D., LEVASSEUR, K., THOMAS, E., TAMAYO, P., DONG, C., ZHOU, Y., LANDER, E. S., NDIAYE, D., WIRTH, D., WINZELER, E. A., MESIROV, J. P. & REGEV, A. 2007. Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients. *Nature*, 450, 1091-1095.
- DALBY, A. R. 2009. A comparative proteomic analysis of the simple amino acid repeat distributions in Plasmodia reveals lineage specific amino acid selection. *PLoS ONE*, 4, e6231.
- DE SILVA, E. K., GEHRKE, A. R., OLSZEWSKI, K., LEON, I., CHAHAL, J. S., BULYK, M. L. & LLINAS, M. 2008. Specific DNA-Binding by Apicomplexan AP2 Transcription Factors. *Proc Natl Acad Sci U S A*, 105, 8393-8398.
- DEBARRY, J. D. & KISSINGER, J. C. 2011. Jumbled genomes: missing Apicomplexan synteny. *Mol Biol Evol*, 28, 2855-71.
- DECHERING, K. J., CUELENAERE, K., KONINGS, R. N. & LEUNISSEN, J. A. 1998. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res*, 26, 4056-62.
- DEITSCH, K., DURAISINGH, M., DZIKOWSKI, R., GUNASEKERA, A., KHAN, S., LE ROCH, K., LLINAS, M., MAIR, G., MCGOVERN, V., ROOS, D., SHOCK, J., SIMS, J., WIEGAND, R. & WINZELER, E. 2007. Mechanisms of gene regulation in *Plasmodium*. *American Journal of Tropical Medicine and Hygiene*, 77, 201-208.
- DEITSCH, K. & DZIKOWSKI, R. 2013. Regulation of gene expression. In *Malaria parasites: comparative genomics, evolution and molecular biology*. Eds Carlton, J., Perkins, S.L., Deitsch, K.W. , Caister Academic Press.
- DEITSCH, K. W. & HVIID, L. 2004. Variant surface antigens, virulence genes and the pathogenesis of malaria. *Trends Parasitol*, 20, 562-6.

- DHARIA, N. V., BRIGHT, A. T., WESTENBERGER, S. J., BARNES, S. W., BATALOV, S., KUHEN, K., BORBOA, R., FEDERE, G. C., MCCLEAN, C. M., VINETZ, J. M., NEYRA, V., LLANOS-CUENTAS, A., BARNWELL, J. W., WALKER, J. R. & WINZELER, E. A. 2010. Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes. *Proc Natl Acad Sci U S A*, 107, 20045-50.
- DONDORP, A. M., NOSTEN, F., YI, P., DAS, D., PHYO, A. P., TARNING, J., LWIN, K. M., ARIEY, F., HANPITHAKPONG, W., LEE, S. J., RINGWALD, P., SILAMUT, K., IMWONG, M., CHOTIVANICH, K., LIM, P., HERDMAN, T., AN, S. S., YEUNG, S., SINGHASIVANON, P., DAY, N. P., LINDEGARDH, N., SOCHEAT, D. & WHITE, N. J. 2009. Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med*, 361, 455-67.
- DUFFY, M. F., SELVARAJAH, S. A., JOSLING, G. A. & PETTER, M. 2012. The role of chromatin in *Plasmodium* gene expression. *Cell Microbiol*, 14, 819-28.
- DURASINGH, M. T., VOSS, T. S., MARTY, A. J., DUFFY, M. F., GOOD, R. T., THOMPSON, J. K., FREITAS-JUNIOR, L. H., SCHERF, A., CRABB, B. S. & COWMAN, A. F. 2005. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell*, 121, 13-24.
- DZIKOWSKI, R., TEMPLETON, T. J. & DEITSCH, K. 2006. Variant antigen gene expression in malaria. *Cellular Microbiology*, 8, 1371-1381.
- EDGAR, A. J. 2003. The gene structure and expression of human ABHD1: overlapping polyadenylation signal sequence with Sec12. *BMC Genomics*, 4, 18.
- ELEMENTO, O., SLONIM, N. & TAVAZOIE, S. 2007. A universal framework for regulatory element discovery across all Genomes and data types. *Molecular Cell*, 28, 337-350.
- EPSTEIN, J. E. & RICHIE, T. L. 2013. The whole parasite, pre-erythrocytic stage approach to malaria vaccine development: a review. *Curr Opin Infect Dis*, 26, 420-8.
- ESHAR, S., DAHAN-PASTERNAK, N., WEINER, A. & DZIKOWSKI, R. 2011. High resolution 3D perspective of *Plasmodium* biology: advancing into a new era. *Trends Parasitol*, 27, 548-54.
- ESSIEN, K., HANNENHALLI, S. & STOECKERT, C. J. 2008. Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to *Plasmodium* phenotypic diversity. *PLoS ONE*, 3, e3122.
- FIELD, Y., KAPLAN, N., FONDUFE-MITTENDORF, Y., MOORE, I. K., SHARON, E., LUBLING, Y., WIDOM, J. & SEGAL, E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol*, 4, e1000216.
- FIGUEIREDO, L. M., FREITAS-JUNIOR, L. H., BOTTIUS, E., OLIVO-MARIN, J. C. & SCHERF, A. 2002. A Central Role for *Plasmodium falciparum* Subtelomeric Regions in Spatial Positioning and Telomere Length Regulation. *Embo J*, 21, 815-24.
- FILION, G. J., VAN BEMMEL, J. G., BRAUNSCHWEIG, U., TALHOUT, W., KIND, J., WARD, L. D., BRUGMAN, W., DE CASTRO, I. J., KERKHOVEN, R. M., BUSSEMAKER, H. J. & VAN STEENSEL, B. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell*, 143, 212-24.
- FLANNERY, E. L., CHATTERJEE, A. K. & WINZELER, E. A. 2013. Antimalarial drug discovery - approaches and progress towards new medicines. *Nat Rev Microbiol*, 11, 849-62.
- FLUECK, C., BARTFAI, R., NIEDERWIESER, I., WITMER, K., ALAKO, B. T., MOES, S., BOZDECH, Z., JENOE, P., STUNNENBERG, H. G. & VOSS, T. S. 2010. A major role for the *Plasmodium falciparum* ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathog*, 6, e1000784.
- FLUECK, C., BARTFAI, R., VOLZ, J., NIEDERWIESER, I., SALCEDO-AMAYA, A. M., ALAKO, B. T., EHLGEN, F., RALPH, S. A., COWMAN, A. F., BOZDECH, Z., STUNNENBERG, H. G. & VOSS, T. S. 2009. *Plasmodium falciparum*

- heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathog*, 5, e1000569.
- FOTH, B. J., ZHANG, N., CHAAL, B. K., SZE, S. K., PREISER, P. R. & BOZDECH, Z. 2011. Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite *Plasmodium falciparum*. *Mol Cell Proteomics*, 10, M110006411.
- FOTH, B. J., ZHANG, N., MOK, S., PREISER, P. R. & BOZDECH, Z. 2008. Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites. *Genome Biol*, 9, R177.
- FRANCIS, S. E., MALKOV, V. A., OLEINIKOV, A. V., ROSSNAGLE, E., WENDLER, J. P., MUTABINGWA, T. K., FRIED, M. & DUFFY, P. E. 2007. Six genes are preferentially transcribed by the circulating and sequestered forms of *Plasmodium falciparum* parasites that infect pregnant women. *Infect Immun*, 75, 4838-50.
- FREESE, J. A., SHARP, B. L., RIDL, F. C. & MARKUS, M. B. 1988. *In vitro* Cultivation of Southern African Strains of *Plasmodium falciparum* and Gametocytogenesis. *S Afr Med J*, 73, 720-2.
- FREITAS-JUNIOR, L. H., BOTTIUS, E., PIRRIT, L. A., DEITSCH, K. W., SCHEIDIG, C., GUINET, F., NEHRBASS, U., WELLEMS, T. E. & SCHERF, A. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*, 407, 1018-22.
- FREITAS-JUNIOR, L. H., HERNANDEZ-RIVAS, R., RALPH, S. A., MONTIEL-CONDADO, D., RUVALCABA-SALAZAR, O. K., ROJAS-MEZA, A. P., MANCIO-SILVA, L., LEAL-SILVESTRE, R. J., GONTIJO, A. M., SHORTE, S. & SCHERF, A. 2005. Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell*, 121, 25-36.
- FREVERT, U., ENGELMANN, S., ZOUGBEDE, S., STANGE, J., NG, B., MATUSCHEWSKI, K., LIEBES, L. & YEE, H. 2005. Intravital observation of *Plasmodium berghei* sporozoite infection of the liver. *PLoS Biol*, 3, e192.
- FRUGIER, M., BOUR, T., AYACH, M., SANTOS, M. A., RUDINGER-THIRION, J., THEOBALD-DIETRICH, A. & PIZZI, E. 2010. Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. *FEBS Lett*, 584, 448-54.
- GANESAN, K., PONMEE, N., JIANG, L., FOWBLE, J. W., WHITE, J., KAMCHONWONGPAISAN, S., YUTHAVONG, Y., WILAIRAT, P. & RATHOD, P. K. 2008. A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates. *PLoS Pathog*, 4, e1000214.
- GARCIA, J. E., PUENTES, A. & PATARROYO, M. E. 2006. Developmental biology of sporozoite-host interactions in *Plasmodium falciparum* malaria: implications for vaccine design. *Clin Microbiol Rev*, 19, 686-707.
- GARDNER, M. J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R. W., CARLTON, J. M., PAIN, A., NELSON, K. E., BOWMAN, S., PAULSEN, I. T., JAMES, K., EISEN, J. A., RUTHERFORD, K., SALZBERG, S. L., CRAIG, A., KYES, S., CHAN, M. S., NENE, V., SHALLOM, S. J., SUH, B., PETERSON, J., ANGIUOLI, S., PERTEA, M., ALLEN, J., SELENGUT, J., HAFT, D., MATHER, M. W., VAIDYA, A. B., MARTIN, D. M. A., FAIRLAMB, A. H., FRAUNHOLZ, M. J., ROOS, D. S., RALPH, S. A., MCFADDEN, G. I., CUMMINGS, L. M., SUBRAMANIAN, G. M., MUNGALL, C., VENTER, J. C., CARUCCI, D. J., HOFFMAN, S. L., NEWBOLD, C., DAVIS, R. W., FRASER, C. M. & BARRELL, B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419, 498-511.
- GOLIGHTLY, L. M., MBACHAM, W., DAILY, J. & WIRTH, D. F. 2000. 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum*. *Mol Biochem Parasitol*, 105, 61-70.

- GONZALES, J. M., PATEL, J. J., PONMEE, N., JIANG, L., TAN, A., MAHER, S. P., WUCHTY, S., RATHOD, P. K. & FERDIG, M. T. 2008. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol*, 6, e238.
- GOPALAKRISHNAN, A. M., NYINDODO, L. A., ROSS FERGUS, M. & LOPEZ-ESTRANO, C. 2009. *Plasmodium falciparum*: preinitiation complex occupancy of active and inactive promoters during erythrocytic stage. *Exp Parasitol*, 121, 46-54.
- GOYAL, M., ALAM, A., IQBAL, M. S., DEY, S., BINDU, S., PAL, C., BANERJEE, A., CHAKRABARTI, S. & BANDYOPADHYAY, U. 2012. Identification and molecular characterization of an Alba-family protein from human malaria parasite *Plasmodium falciparum*. *Nucleic Acids Res*, 40, 1174-90.
- GREGORY, T. R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc*, 76, 65-101.
- GUINET, F., DVORAK, J. A., FUJIOKA, H., KEISTER, D. B., MURATOVA, O., KASLOW, D. C., AIKAWA, M., VAIDYA, A. B. & WELLEMS, T. E. 1996. A Developmental Defect in *Plasmodium falciparum* Male Gametogenesis. *J. Cell Biol.*, 135, 269-278.
- GUNASEKERA, A. M., MYRICK, A., MILITELLO, K. T., SIMS, J. S., DONG, C. K., GIERAHN, T., LE ROCH, K., WINZELER, E. & WIRTH, D. F. 2007. Regulatory motifs uncovered among gene expression clusters in *Plasmodium falciparum*. *Mol Biochem Parasitol*, 153, 19-30.
- GUNASEKERA, A. M., PATANKAR, S., SCHUG, J., EISEN, G., KISSINGER, J., ROOS, D. & WIRTH, D. F. 2004. Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Mol Biochem Parasitol*, 136, 35-42.
- GUNASEKERA, A. M., PATANKAR, S., SCHUG, J., EISEN, G. & WIRTH, D. F. 2003. Drug-induced alterations in gene expression of the asexual blood forms of *Plasmodium falciparum*. *Molecular Microbiology*, 50, 1229-1239.
- GUPTA, A. P., CHIN, W. H., ZHU, L., MOK, S., LUAH, Y. H., LIM, E. H. & BOZDECH, Z. 2013. Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite *Plasmodium falciparum*. *PLoS Pathog*, 9, e1003170.
- HALL, N., KARRAS, M., RAINE, J. D., CARLTON, J. M., KOUIJ, T. W., BERRIMAN, M., FLORENS, L., JANSSEN, C. S., PAIN, A., CHRISTOPHIDES, G. K., JAMES, K., RUTHERFORD, K., HARRIS, B., HARRIS, D., CHURCHER, C., QUAIL, M. A., ORMOND, D., DOGGETT, J., TRUEMAN, H. E., MENDOZA, J., BIDWELL, S. L., RAJANDREAM, M. A., CARUCCI, D. J., YATES, J. R., 3RD, KAFATOS, F. C., JANSE, C. J., BARRELL, B., TURNER, C. M., WATERS, A. P. & SINDEN, R. E. 2005. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 307, 82-6.
- HASENKAMP, S., RUSSELL, K. T. & HORROCKS, P. 2012a. Comparison of the absolute and relative efficiencies of electroporation-based transfection protocols for *Plasmodium falciparum*. *Malar J*, 11, 210.
- HASENKAMP, S., WONG, E. H. & HORROCKS, P. 2012b. An improved single-step lysis protocol to measure luciferase bioluminescence in *Plasmodium falciparum*. *Malar J*, 11, 42.
- HERMSEN, R., TEN WOLDE, P. R. & TEICHMANN, S. 2008. Chance and necessity in chromosomal gene distributions. *Trends Genet*, 24, 216-9.
- HERNANDEZ-RIVAS, R., HERRERA-SOLORIO, A. M., SIERRA-MIRANDA, M., DELGADILLO, D. M. & VARGAS, M. 2013. Impact of chromosome ends on the biology and virulence of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 187, 121-8.
- HERNANDEZ-RIVAS, R., PEREZ-TOLEDO, K., HERRERA SOLORIO, A. M., DELGADILLO, D. M. & VARGAS, M. 2010. Telomeric heterochromatin in *Plasmodium falciparum*. *J Biomed Biotechnol*, 2010, 290501.
- HILLS, T., SRIVASTAVA, A., AYI, K., WERNIMONT, A. K., KAIN, K., WATERS, A. P., HUI, R. & PIZARRO, J. C. 2011. Characterization of a new phosphatase from *Plasmodium*. *Mol Biochem Parasitol*, 179, 69-79.
- HIROSE, S., PAYNE, S. H. & LOOMIS, W. F. 2006. cis-Acting site controlling bidirectional transcription at the growth-differentiation transition in *Dictyostelium discoideum*. *Eukaryot Cell*, 5, 1104-10.

- HO, M. R., TSAI, K. W. & LIN, W. C. 2012. A unified framework of overlapping genes: towards the origination and endogenic regulation. *Genomics*, 100, 231-9.
- HOEIJMAKERS, W. A., FLUECK, C., FRANCOIJS, K. J., SMITS, A. H., WETZEL, J., VOLZ, J. C., COWMAN, A. F., VOSS, T., STUNNENBERG, H. G. & BARTFAI, R. 2012a. *Plasmodium falciparum* centromeres display a unique epigenetic makeup and cluster prior to and during schizogony. *Cell Microbiol*, 14, 1391-401.
- HOEIJMAKERS, W. A., SALCEDO-AMAYA, A. M., SMITS, A. H., FRANCOIJS, K. J., TREECK, M., GILBERGER, T. W., STUNNENBERG, H. G. & BARTFAI, R. 2013. H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the *Plasmodium falciparum* genome. *Mol Microbiol*, 87, 1061-73.
- HOEIJMAKERS, W. A., STUNNENBERG, H. G. & BARTFAI, R. 2012b. Placing the *Plasmodium falciparum* epigenome on the map. *Trends Parasitol*, 28, 486-95.
- HORROCKS, P., DECHERING, K. & LANZER, M. 1998. Control of gene expression in *Plasmodium falciparum*. *Mol Biochem Parasitol*, 95, 171-81.
- HORROCKS, P. & KILBEY, B. J. 1996. Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. *Mol Biochem Parasitol*, 82, 207-215.
- HORROCKS, P. & LANZER, M. 1999. Differences in nucleosome organization over episomally located plasmids coincides with aberrant promoter activity in *P. falciparum*. *Parasitology International*, 48, 55-61.
- HORROCKS, P., WONG, E., RUSSELL, K. & EMES, R. D. 2009. Control of gene expression in *Plasmodium falciparum* - ten years on. *Mol Biochem Parasitol*, 164, 9-25.
- HU, G., CABRERA, A., KONO, M., MOK, S., CHAAL, B. K., HAASE, S., ENGELBERG, K., CHEEMADAN, S., SPIELMANN, T., PREISER, P. R., GILBERGER, T. W. & BOZDECH, Z. 2010. Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nat Biotechnol*, 28, 91-8.
- JANSE, C. J., VAN DER KLOOSTER, P. F., VAN DER KAAY, H. J., VAN DER PLOEG, M. & OVERDULVE, J. P. 1986. DNA synthesis in *Plasmodium berghei* during asexual and sexual development. *Mol Biochem Parasitol*, 20, 173-82.
- JOANNIN, N., ABHIMAN, S., SONNHAMMER, E. L. & WAHLGREN, M. 2008. Sub-grouping and sub-functionalization of the RIFIN multi-copy protein family. *BMC Genomics*, 9, 19.
- JOHNSON, Z. I. & CHISHOLM, S. W. 2004. Properties of overlapping genes are conserved across microbial genomes. *Genome Res*, 14, 2268-72.
- JURGELENAITE, R., DIJKSTRA, T. M., KOCKEN, C. H. & HESKES, T. 2009. Gene regulation in the intraerythrocytic cycle of *Plasmodium falciparum*. *Bioinformatics*, 25, 1484-91.
- KORBER, P. 2012. Active nucleosome positioning beyond intrinsic biophysics is revealed by in vitro reconstitution. *Biochem Soc Trans*, 40, 377-82.
- KUEHN, A. & PRADEL, G. 2010. The coming-out of malaria gametocytes. *J Biomed Biotechnol*, 2010, 976827.
- KYES, S., CHRISTODOULOU, Z., PINCHES, R. & NEWBOLD, C. 2002. Stage-specific merozoite surface protein 2 antisense transcripts in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, 123, 79-83.
- KYES, S., PINCHES, R. & NEWBOLD, C. 2000. A simple RNA analysis method shows *var* and *rif* multigene family expression patterns in *Plasmodium falciparum*. *Molecular and Biochemical Parasitology*, 105, 311-315.
- KYES, S. A., ROWE, J. A., KRIEK, N. & NEWBOLD, C. I. 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, 9333-8.
- LALLOO, D. G., SHINGADIA, D., PASVOL, G., CHIODINI, P. L., WHITTY, C. J., BEECHING, N. J., HILL, D. R., WARRELL, D. A., BANNISTER, B. A. & TRAVELLERS, H. P. A. A. C. O. M. P. I. U. 2007. UK malaria treatment guidelines. *J Infect*, 54, 111-21.
- LAMBROS, C. & VANDERBERG, J. P. 1979. Synchronization of *Plasmodium falciparum* Erythrocytic Stages in Culture. *J Parasitol*, 65, 418-20.

- LANZER, M., DE BRUIN, D. & RAVETCH, J. V. 1992a. Transcription Mapping of a 100 kb Locus of *Plasmodium falciparum* Identifies an Intergenic Region in Which Transcription Terminates and Reinitiates. *Embo J*, 11, 1949-55.
- LANZER, M., DE BRUIN, D. & RAVETCH, J. V. 1992b. Transcription mapping of a 100 kb locus of *Plasmodium falciparum* identifies an intergenic region in which transcription terminates and reinitiates. *EMBO Journal*, 11, 1949-55.
- LAVIGNE, M. & BUC, H. 1999. Compression of the DNA minor groove is responsible for termination of DNA synthesis by HIV-1 reverse transcriptase. *J Mol Biol*, 285, 977-95.
- LE ROCH, K. G., JOHNSON, J. R., AHIBOH, H., CHUNG, D. W., PRUDHOMME, J., PLOUFFE, D., HENSON, K., ZHOU, Y., WITOLA, W., YATES, J. R., MAMOUN, C. B., WINZELER, E. A. & VIAL, H. 2008. A systematic approach to understand the mechanism of action of the bisthiazolium compound T4 on the human malaria parasite, *Plasmodium falciparum*. *BMC Genomics*, 9, 513.
- LE ROCH, K. G., JOHNSON, J. R., FLORENS, L., ZHOU, Y., SANTROSYAN, A., GRAINGER, M., YAN, S. F., WILLIAMSON, K. C., HOLDER, A. A., CARUCCI, D. J., YATES, J. R., III & WINZELER, E. A. 2004. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Research*, 14, 2308-2318.
- LE ROCH, K. G., ZHOU, Y. Y., BLAIR, P. L., GRAINGER, M., MOCH, J. K., HAYNES, J. D., DE LA VEGA, P., HOLDER, A. A., BATALOV, S., CARUCCI, D. J. & WINZELER, E. A. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301, 1503-1508.
- LEMIEUX, J. E., GOMEZ-ESCOBAR, N., FELLER, A., CARRET, C., AMAMBUA-NGWA, A., PINCHES, R., DAY, F., KYES, S. A., CONWAY, D. J., HOLMES, C. C. & NEWBOLD, C. I. 2009. Statistical estimation of cell-cycle progression and lineage commitment in *Plasmodium falciparum* reveals a homogeneous pattern of transcription in ex vivo culture. *Proc Natl Acad Sci U S A*, 106, 7559-64.
- LEVINE, M. 2011. Paused RNA polymerase II as a developmental checkpoint. *Cell*, 145, 502-11.
- LEVITT, A. 1993. RNA processing in malarial parasites. *Parasitol Today*, 9, 465-8.
- LI, F., SONBUCHNER, L., KYES, S. A., EPP, C. & DEITSCH, K. W. 2008. Nuclear non-coding RNAs are transcribed from the centromeres of *Plasmodium falciparum* and are associated with centromeric chromatin. *J. Biol. Chem.*, 283, 5692-5698.
- LINDNER, S. E., DE SILVA, E. K., KECK, J. L. & LLINAS, M. 2010. Structural determinants of DNA binding by a *P. falciparum* ApiAP2 transcriptional regulator. *J Mol Biol*, 395, 558-67.
- LLINAS, M., BOZDECH, Z., WONG, E. D., ADAI, A. T. & DERISI, J. L. 2006. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Research*, 34, 1166-73.
- LOPEZ-RUBIO, J. J., GONTIJO, A. M., NUNES, M. C., ISSAR, N., RIVAS, R. H. & SCHERF, A. 2007. 5' flanking region of *var* genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol. Microbiol.*, 66, 1296-1305.
- LOPEZ-RUBIO, J. J., MANCIO-SILVA, L. & SCHERF, A. 2009. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe*, 5, 179-90.
- MACKINNON, M. J., LI, J., MOK, S., KORTOK, M. M., MARSH, K., PREISER, P. R. & BOZDECH, Z. 2009. Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. *PLoS Pathog*, 5, e1000644.
- MAIR, G. R., BRAKS, J. A. M., GARVER, L. S., WIEGANT, L., HALL, N., DIRKS, R. W., KHAN, S. M., DIMOPOULOS, G., JANSE, C. J. & WATERS, A. P. 2006. Regulation of sexual development of *Plasmodium* by translational repression. *Science*, 313, 667-669.
- MAIR, G. R., LASONDER, E., GARVER, L. S., FRANKE-FAYARD, B. M., CARRET, C. K., WIEGANT, J. C., DIRKS, R. W., DIMOPOULOS, G., JANSE, C. J. & WATERS, A. P. 2010. Universal features of post-transcriptional gene regulation are critical for *Plasmodium* zygote development. *PLoS Pathog*, 6, e1000767.

- MAKALOWSKA, I., LIN, C. F. & MAKALOWSKI, W. 2005. Overlapping genes in vertebrate genomes. *Comput Biol Chem*, 29, 1-12.
- MATUSCHEWSKI, K. 2006. Getting infectious: formation and maturation of Plasmodium sporozoites in the Anopheles vector. *Cell Microbiol*, 8, 1547-56.
- MAVRICH, T. N., IOSHIKHES, I. P., VENTERS, B. J., JIANG, C., TOMSHO, L. P., QI, J., SCHUSTER, S. C., ALBERT, I. & PUGH, B. F. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res*, 18, 1073-83.
- MIAO, J., FAN, Q., CUI, L., LI, J., LI, J. & CUI, L. 2006. The malaria parasite *Plasmodium falciparum* histones: organization, expression, and acetylation. *Gene*, 369, 53-65.
- MIAO, J., FAN, Q., CUI, L., LI, X., WANG, H., NING, G., REESE, J. C. & CUI, L. 2010. The MYST family histone acetyltransferase regulates gene expression and cell cycle in malaria parasite *Plasmodium falciparum*. *Mol Microbiol*, 78, 883-902.
- MILITELLO, K. T., DODGE, M., BETHKE, L. & WIRTH, D. F. 2004. Identification of regulatory elements in the *Plasmodium falciparum* genome. *Molecular and Biochemical Parasitology*, 134, 75-88.
- MILITELLO, K. T., PATEL, V., CHESSLER, A. D., FISHER, J. K., KASPER, J. M., GUNASEKERA, A. & WIRTH, D. F. 2005. RNA polymerase II synthesizes antisense RNA in *Plasmodium falciparum*. *RNA - Pub.RNA Soc.*, 11, 365-370.
- MILITELLO, K. T., REFOUR, P., COMEAUX, C. A. & DURAISINGH, M. T. 2008. Antisense RNA and RNAi in protozoan parasites: Working hard or hardly working? *Mol. Biochem. Parasitol.*, 157, 117-126.
- MILNER, D. A., JR., POCHET, N., KRUPKA, M., WILLIAMS, C., SEYDEL, K., TAYLOR, T. E., VAN DE PEER, Y., REGEV, A., WIRTH, D., DAILY, J. P. & MESIROV, J. P. 2012. Transcriptional profiling of *Plasmodium falciparum* parasites from patients with severe malaria identifies distinct low vs. high parasitemic clusters. *PLoS One*, 7, e40739.
- MIRSKY, A. E. & RIS, H. 1951. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J Gen Physiol*, 34, 451-62.
- MOK, S., IMWONG, M., MACKINNON, M. J., SIM, J., RAMADOSS, R., YI, P., MAYXAY, M., CHOTIVANICH, K., LIONG, K. Y., RUSSELL, B., SOCHEAT, D., NEWTON, P. N., DAY, N. P., WHITE, N. J., PREISER, P. R., NOSTEN, F., DONDORP, A. M. & BOZDECH, Z. 2011. Artemisinin resistance in *Plasmodium falciparum* is associated with an altered temporal pattern of transcription. *BMC Genomics*, 12, 391.
- MOTA, M. M. & RODRIGUEZ, A. 2004. Migration through host cells: the first steps of *Plasmodium* sporozoites in the mammalian host. *Cell Microbiol*, 6, 1113-8.
- MOURIER, T., CARRET, C., KYES, S., CHRISTODOULOU, Z., GARDNER, P. P., JEFFARES, D. C., PINCHES, R., BARRELL, B., BERRIMAN, M., GRIFFITHS-JONES, S., IVENS, A., NEWBOLD, C. & PAIN, A. 2008. Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*. *Gen. Research*, 18, 281-292.
- MURRAY, C. J., ROSENFELD, L. C., LIM, S. S., ANDREWS, K. G., FOREMAN, K. J., HARING, D., FULLMAN, N., NAGHAVI, M., LOZANO, R. & LOPEZ, A. D. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet*, 379, 413-31.
- MYRICK, A., SARR, O., DIENG, T., NDIR, O., MBOUP, S. & WIRTH, D. F. 2005. Analysis of the genetic diversity of the *Plasmodium falciparum* multidrug resistance gene 5' upstream region. *Am J Trop Med Hyg*, 72, 182-8.
- NATALANG, O., BISCHOFF, E., DEPLAINE, G., PROUX, C., DILLIES, A., SISMEIRO, O., GUIGON, G., BONNEFOY, S., PATARAPOTIKUL, J., MERCEREAU-PUIJALON, O., COPPEE, J. & DAVID, P. H. 2008. Dynamic RNA profiling in *Plasmodium falciparum* synchronized blood stages exposed to lethal doses of artesunate. *BMC Genomics*, [in press].
- NEAFSEY, D. E., HARTL, D. L. & BERRIMAN, M. 2005. Evolution of noncoding and silent coding sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes. *Mol Biol Evol*, 22, 1621-6.

- NEIL, H., MALABAT, C., D'AUBENTON-CARAFI, Y., XU, Z., STEINMETZ, L. M. & JACQUIER, A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457, 1038-42.
- NKRUMAH, L. J., MUHLE, R. A., MOURA, P. A., GHOSH, P., HATFULL, G. F., JACOBS, W. R. & FIDOCK, D. A. 2006. Efficient Site-Specific Integration in *Plasmodium falciparum* Chromosomes Mediated by Mycobacteriophage *Bxb1* Integrase. *Nat Meth*, 3, 615-621.
- NYGAARD, S., BRAUNSTEIN, A., MALSEN, G., VAN DONGEN, S., GARDNER, P. P., KROGH, A., OTTO, T. D., PAIN, A., BERRIMAN, M., MCAULIFFE, J., DERMITZAKIS, E. T. & JEFFARES, D. C. 2010. Long- and short-term selective forces on malaria parasite genomes. *PLoS Genet*, 6.
- OHINATA, Y., SUTOU, S., KONDO, M., TAKAHASHI, T. & MITSUI, Y. 2002. Male-enhanced antigen-1 gene flanked by two overlapping genes is expressed in late spermatogenesis. *Biol Reprod*, 67, 1824-31.
- OSATO, N., SUZUKI, Y., IKEO, K. & GOJOBORI, T. 2007. Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics*, 176, 1299-306.
- OTTO, T. D., WILINSKI, D., ASSEFA, S., KEANE, T. M., SARRY, L. R., BOHME, U., LEMIEUX, J., BARRELL, B., PAIN, A., BERRIMAN, M., NEWBOLD, C. & LLINAS, M. 2010. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol Microbiol*, 76, 12-24.
- PAIN, A., BOHME, U., BERRY, A. E., MUNGALL, K., FINN, R. D., JACKSON, A. P., MOURIER, T., MISTRY, J., PASINI, E. M., ASLETT, M. A., BALASUBRAMMANIAM, S., BORGHARDT, K., BROOKS, K., CARRET, C., CARVER, T. J., CHEREVACH, I., CHILLINGWORTH, T., CLARK, T. G., GALINSKI, M. R., HALL, N., HARPER, D., HARRIS, D., HAUSER, H., IVENS, A., JANSSEN, C. S., KEANE, T., LARKE, N., LAPP, S., MARTI, M., MOULE, S., MEYER, I. M., ORMOND, D., PETERS, N., SANDERS, M., SANDERS, S., SARGEANT, T. J., SIMMONDS, M., SMITH, F., SQUARES, R., THURSTON, S., TIVEY, A. R., WALKER, D., WHITE, B., ZUIDERWIJK, E., CHURCHER, C., QUAIL, M. A., COWMAN, A. F., TURNER, C. M., RAJANDREAM, M. A., KOCKEN, C. H., THOMAS, A. W., NEWBOLD, C. I., BARRELL, B. G. & BERRIMAN, M. 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, 455, 799-803.
- PAINTER, H. J., CAMPBELL, T. L. & LLINAS, M. 2011. The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. *Mol Biochem Parasitol*, 176, 1-7.
- PANCSA, R. & TOMPA, P. 2012. Structural disorder in eukaryotes. *PLoS ONE*, 7, e34687.
- PASTERNAK, N. D. & DZIKOWSKI, R. 2009. PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. *Int J Biochem Cell Biol*, 41, 1463-6.
- PEREZ-TOLEDO, K., ROJAS-MEZA, A. P., MANCIO-SILVA, L., HERNANDEZ-CUEVAS, N. A., DELGADILLO, D. M., VARGAS, M., MARTINEZ-CALVILLO, S., SCHERF, A. & HERNANDEZ-RIVAS, R. 2009. *Plasmodium falciparum* heterochromatin protein 1 binds to tri-methylated histone 3 lysine 9 and is linked to mutually exclusive expression of var genes. *Nucleic Acids Res*, 37, 2596-606.
- PETTER, M., LEE, C. C., BYRNE, T. J., BOYSEN, K. E., VOLZ, J., RALPH, S. A., COWMAN, A. F., BROWN, G. V. & DUFFY, M. F. 2011. Expression of *P. falciparum* var genes involves exchange of the histone variant H2A.Z at the promoter. *PLoS Pathog*, 7, e1001292.
- POLSON, H. E. J. & BLACKMAN, M. J. 2005. A role for poly (dA)poly(dT) tracts in directing activity of the *Plasmodium falciparum* calmodulin gene promoter. *Mol. Biochem. Parasitol.*, 141, 179-189.
- PONTS, N., CHUNG, D. W. & LE ROCH, K. G. 2012. Strand-specific RNA-seq applied to malaria samples. *Methods Mol Biol*, 883, 59-73.
- PONTS, N., FU, L., HARRIS, E. Y., ZHANG, J., CHUNG, D. W., CERVANTES, M. C., PRUDHOMME, J., ATANASOVA-PENICHON, V., ZEHRAOUI, E., BUNNIK, E. M., RODRIGUES, E. M., LONARDI, S., HICKS, G. R., WANG, Y. & LE ROCH, K. G.

2013. Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*. *Cell Host Microbe*, 14, 696-706.
- PONTS, N., HARRIS, E. Y., LONARDI, S. & LE ROCH, K. G. 2011. Nucleosome occupancy at transcription start sites in the human malaria parasite: a hard-wired evolution of virulence? *Infect Genet Evol*, 11, 716-24.
- PONTS, N., HARRIS, E. Y., PRUDHOMME, J., WICK, I., ECKHARDT-LUDKA, C., HICKS, G. R., HARDIMAN, G., LONARDI, S. & LE ROCH, K. G. 2010. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res*, 20, 228-38.
- PORTER, M. E. 2002. Positive and negative effects of deletions and mutations within the 5' flanking sequences of *Plasmodium falciparum* DNA polymerase delta. *Molecular and Biochemical Parasitology*, 122, 9-19.
- RAABE, A. C., BILLKER, O., VIAL, H. J. & WENGELNIK, K. 2009. Quantitative assessment of DNA replication to monitor microgametogenesis in *Plasmodium berghei*. *Mol Biochem Parasitol*, 168, 172-6.
- RALPH, S. A., SCHEIDIG-BENATAR, C. & SCHERF, A. 2005. Antigenic variation in *Plasmodium falciparum* is associated with movement of *var* loci between subnuclear locations. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 5414-5419.
- ROY, S. W. & PENNY, D. 2007. Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. *Mol Biol Evol*, 24, 1926-33.
- RUVALCABA-SALAZAR, O. K., RAMIREZ-ESTUDILLO, M. D. C., MONTIEL-CONDADO, D., RECILLAS-TARGA, F., VARGAS, M. & HERNANDEZ-RIVAS, R. 2005. Recombinant and Native *Plasmodium falciparum* TATA-Binding-Protein Binds to a Specific TATA Box Element in Promoter Regions. *Molecular and Biochemical Parasitology*, 140, 183-196.
- RUVOLO, V., ALTSZULER, R. & LEVITT, A. 1993. The transcript encoding the circumsporozoite antigen of *Plasmodium berghei* utilizes heterogeneous polyadenylation sites. *Mol Biochem Parasitol*, 57, 137-50.
- SANNA, C. R., LI, W. H. & ZHANG, L. 2008. Overlapping genes in the human and mouse genomes. *BMC Genomics*, 9, 169.
- SCHERF, A., FIGUEIREDO, L. M. & FREITAS-JUNIOR, L. H. 2001. *Plasmodium* Telomeres: A Pathogen's Perspective. *Current Opinion in Microbiology*, 4, 409-414.
- SCHERF, A., LOPEZ-RUBIO, J. J. & RIVIERE, L. 2008. Antigenic Variation in *Plasmodium falciparum*. *Annual Review of Microbiology*, 62, 445-470.
- SEGAL, E. & WIDOM, J. 2009a. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet*, 10, 443-56.
- SEGAL, E. & WIDOM, J. 2009b. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol*, 19, 65-71.
- SEGAL, E. & WIDOM, J. 2009c. What controls nucleosome positions? *Trends Genet*, 25, 335-43.
- SHOCK, J. L., FISCHER, K. F. & DERISI, J. L. 2007. Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biology*, 8.
- SIERRA-MIRANDA, M., DELGADILLO, D. M., MANCIO-SILVA, L., VARGAS, M., VILLEGAS-SEPULVEDA, N., MARTINEZ-CALVILLO, S., SCHERF, A. & HERNANDEZ-RIVAS, R. 2012. Two long non-coding RNAs generated from subtelomeric regions accumulate in a novel perinuclear compartment in *Plasmodium falciparum*. *Mol Biochem Parasitol*, 185, 36-47.
- SIMS, J. S., MILITELLO, K. T., SIMS, P. A., PATEL, V. P., KASPER, J. M. & WIRTH, D. F. 2009. Patterns of gene-specific and total transcriptional activity during the *Plasmodium falciparum* intraerythrocytic developmental cycle. *Eukaryot Cell*, 8, 327-38.
- SINGH, B. & DANESHVAR, C. 2013. Human infections and detection of *Plasmodium knowlesi*. *Clin Microbiol Rev*, 26, 165-84.

- SORBER, K., DIMON, M. T. & DERISI, J. L. 2011. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res*, 39, 3820-35.
- STILLMAN, B. 1996. Comparison of DNA replication in cells from Prokarya and Eukarya. In DNA replication in Eukaryotic Cells. Cold Spring Harbor Laboratory Press, USA.
- STURM, A., AMINO, R., VAN DE SAND, C., REGEN, T., RETZLAFF, S., RENNENBERG, A., KRUEGER, A., POLLOK, J. M., MENARD, R. & HEUSSLER, V. T. 2006. Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science*, 313, 1287-90.
- SZAFRANSKI, K., LEHMANN, R., PARRA, G., GUIGO, R. & GLOCKNER, G. 2005. Gene organization features in A/T-rich organisms. *J Mol Evol*, 60, 90-8.
- TARUN, A. S., BAER, K., DUMPIT, R. F., GRAY, S., LEJARCEGUI, N., FREVERT, U. & KAPPE, S. H. 2006. Quantitative isolation and in vivo imaging of malaria parasite liver stages. *Int J Parasitol*, 36, 1283-93.
- TEMPLETON, T. J. 2009. The varieties of gene amplification, diversification and hypervariability in the human malaria parasite, *Plasmodium falciparum*. *Mol Biochem Parasitol*, 166, 109-16.
- TEUSCHER, F., GATTON, M. L., CHEN, N., PETERS, J., KYLE, D. E. & CHENG, Q. 2010. Artemisinin-induced dormancy in *plasmodium falciparum*: duration, recovery rates, and implications in treatment failure. *J Infect Dis*, 202, 1362-8.
- THOMAS, C. A., JR. 1971. The genetic organization of chromosomes. *Annu Rev Genet*, 5, 237-56.
- TILLEY, L., DIXON, M. W. & KIRK, K. 2011. The *Plasmodium falciparum*-infected red blood cell. *Int J Biochem Cell Biol*, 43, 839-42.
- TIROSH, I. & BARKAI, N. 2008. Two strategies for gene regulation by promoter nucleosomes. *Genome Res*, 18, 1084-91.
- TIROSH, I., BERMAN, J. & BARKAI, N. 2007. The pattern and evolution of yeast promoter bendability. *Trends Genet*, 23, 318-21.
- TRAGER, W. & JENSEN, J. B. 1976. Human Malaria Parasites in Continuous Culture. *Science*, 193, 673-5.
- TRELLE, M. B., SALCEDO-AMAYA, A. M., COHEN, A. M., STUNNENBERG, H. G. & JENSEN, O. N. 2009. Global histone analysis by mass spectrometry reveals a high content of acetylated lysine residues in the malaria parasite *Plasmodium falciparum*. *J Proteome Res*, 8, 3439-50.
- TUIKUE NDAM, N., BISCHOFF, E., PROUX, C., LAVSTSEN, T., SALANTI, A., GUITARD, J., NIELSEN, M. A., COPPEE, J. Y., GAYE, A., THEANDER, T., DAVID, P. H. & DELORON, P. 2008. *Plasmodium falciparum* transcriptome analysis reveals pregnancy malaria associated gene expression. *PLoS ONE*, 3, e1855.
- ULLU, E., TSCHUDI, C. & CHAKRABORTY, T. 2004. RNA interference in protozoan parasites. *Cellular Microbiology*, 6, 509-519.
- UPADHYAY, R., BAWANKAR, P., MALHOTRA, D. & PATANKAR, S. 2005. A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium falciparum*. *Molecular and Biochemical Parasitology*, 144, 149-158.
- VAN NOORT, V. & HUYNEN, M. A. 2006. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends in Genetics*, 22, 73-78.
- VANDERBERG, J. P. 1977. *Plasmodium berghei*: quantitation of sporozoites injected by mosquitoes feeding on a rodent host. *Exp Parasitol*, 42, 169-81.
- VAUGHAN, A. M., ALY, A. S. & KAPPE, S. H. 2008. Malaria parasite pre-erythrocytic stage infection: gliding and hiding. *Cell Host Microbe*, 4, 209-18.
- VAUGHAN, A. M., WANG, R. & KAPPE, S. H. 2010. Genetically engineered, attenuated whole-cell vaccine approaches for malaria. *Hum Vaccin*, 6, 107-13.
- VEERAMACHANENI, V., MAKALOWSKI, W., GALDZICKI, M., SOOD, R. & MAKALOWSKA, I. 2004. Mammalian overlapping genes: the comparative perspective. *Genome Res*, 14, 280-6.
- VIALLI, M. 1957. [Desoxyribonucleic acid volume & content per nucleus]. *Exp Cell Res*, 13, 284-93.

- VIGNALI, M., ARMOUR, C. D., CHEN, J., MORRISON, R., CASTLE, J. C., BIERY, M. C., BOUZEK, H., MOON, W., BABAK, T., FRIED, M., RAYMOND, C. K. & DUFFY, P. E. 2011. NSR-seq transcriptional profiling enables identification of a gene signature of *Plasmodium falciparum* parasites infecting children. *J Clin Invest*, 121, 1119-29.
- VOLZ, J., CARVALHO, T. G., RALPH, S. A., GILSON, P., THOMPSON, J., TONKIN, C. J., LANGER, C., CRABB, B. S. & COWMAN, A. F. 2010. Potential epigenetic regulatory proteins localise to distinct nuclear sub-compartments in *Plasmodium falciparum*. *Int J Parasitol*, 40, 109-21.
- VOLZ, J. C., BARTFAI, R., PETTER, M., LANGER, C., JOSLING, G. A., TSUBOI, T., SCHWACH, F., BAUM, J., RAYNER, J. C., STUNNENBERG, H. G., DUFFY, M. F. & COWMAN, A. F. 2012. PfSET10, a *Plasmodium falciparum* methyltransferase, maintains the active var gene in a poised state during parasite division. *Cell Host Microbe*, 11, 7-18.
- VOSS, T. S., TONKIN, C. J., MARTY, A. J., THOMPSON, J. K., HEALER, J., CRABB, B. S. & COWMAN, A. F. 2007. Alterations in Local Chromatin Environment are Involved in Silencing and Activation of Subtelomeric var Genes in *Plasmodium falciparum*. *Molecular Microbiology*, 66, 139-150.
- WATANABE, J., SASAKI, M., SUZUKI, Y. & SUGANO, S. 2001. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*. *Nucleic Acids Res*, 29, 70-1.
- WATANABE, J., SASAKI, M., SUZUKI, Y. & SUGANO, S. 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, 105-13.
- WATANABE, J., SUZUKI, Y., SASAKI, M. & SUGANO, S. 2004. Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites. *Nucleic Acids Res*, 32, D334-8.
- WATANABE, J., WAKAGURI, H., SASAKI, M., SUZUKI, Y. & SUGANO, S. 2007. Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs. *Nucleic Acids Res*, 35, D431-8.
- WEINER, A., DAHAN-PASTERNAK, N., SHIMONI, E., SHINDER, V., VON HUTH, P., ELBAUM, M. & DZIKOWSKI, R. 2011. 3D nuclear architecture reveals coupled cell cycle dynamics of chromatin and nuclear pores in the malaria parasite *Plasmodium falciparum*. *Cell Microbiol*, 13, 967-77.
- WESTENBERGER, S. J., CUI, L., DHARIA, N., WINZELER, E. & CUI, L. 2009. Genome-wide nucleosome mapping of *Plasmodium falciparum* reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes. *BMC Genomics*, 10, 610.
- WHITE, N. J., PUKRITTAYAKAMEE, S., HIEN, T. T., FAIZ, M. A., MOKUOLU, O. A. & DONDORP, A. M. 2014. Malaria. *Lancet*, 383, 723-35.
- WHO 2013. World Malaria Report 2013. World Health Organisation.
- WONG, E. H., HASENKAMP, S. & HORROCKS, P. 2011. Analysis of the molecular mechanisms governing the stage-specific expression of a prototypical housekeeping gene during intraerythrocytic development of *P. falciparum*. *J Mol Biol*, 408, 205-21.
- WOODS, K. K., MAEHIGASHI, T., HOWERTON, S. B., SINES, C. C., TANNENBAUM, S. & WILLIAMS, L. D. 2004. High-resolution structure of an extended A-tract: [d(CGCAAATTTGCG)]₂. *J Am Chem Soc*, 126, 15330-1.
- WU, J., SIEGLAFF, D. H., GERVIN, J. & XIE, X. S. 2008. Discovering regulatory motifs in the *Plasmodium* genome using comparative genomics. *Bioinformatics*, 24, 1843-1849.
- WU, R. & LI, H. 2010. Positioned and G/C-capped poly(dA:dT) tracts associate with the centers of nucleosome-free regions in yeast promoters. *Genome Res*, 20, 473-84.
- YADAV, M. K. & SWATI, D. 2012. Comparative genome analysis of six malarial parasites using codon usage bias based tools. *Bioinformation*, 8, 1230-9.

- YOUNG, J., JOHNSON, J., BENNER, C., YAN, S. F., CHEN, K., LE ROCH, K., ZHOU, Y. & WINZELER, E. 2008. In Silico Discovery of Transcription Regulatory Elements in *Plasmodium falciparum*. *BMC Genomics*, 9, 70.
- ZANOTTO, E., HAKKINEN, A., TEKU, G., SHEN, B., RIBEIRO, A. S. & JACOBS, H. T. 2009. NF-Y influences directionality of transcription from the bidirectional Mrps12/Sarsm promoter in both mouse and human cells. *Biochim Biophys Acta*, 1789, 432-42.
- ZHOU, C. & BLUMBERG, B. 2003. Overlapping gene structure of human VLCAD and DLG4. *Gene*, 305, 161-6.
- ZHOU, Y., BIZZARO, J. W. & MARX, K. A. 2004. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. *BMC Genomics*, 5, 95.
- ZILVERSMIT, M. M., VOLKMAN, S. K., DEPRISTO, M. A., WIRTH, D. F., AWADALLA, P. & HARTL, D. L. 2010. Low-complexity regions in *Plasmodium falciparum*: missing links in the evolution of an extreme genome. *Mol Biol Evol*, 27, 2198-209.

<http://www.standard.co.uk/news/health/scientists-raise-hope-of-new-lifesaving-malaria-vaccine-8911039.html>

http://www.who.int/malaria/publications/world_malaria_report_2013/report/en/