Analysis of transcription of the *Plasmodium interspersed repeat (pir)* multigene family across the life cycle of *Plasmodium* species

Timothy Stephen Little

Department of Life Sciences, Imperial College London

and

The Francis Crick Institute

PhD Supervisor: Jean Langhorne and George Christophides

A thesis submitted for the degree of

Doctor of Philosophy

Imperial College

October 2022

Declaration

I Timothy Stephen Little confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Part of this thesis has been published previously as:

Little, T. S., Cunningham, D. A., Vandomme, A., Lopez, C. T., Amis, S., Alder, C., Addy, J. W. G., McLaughlin, S., Hosking, C., Christophides, G., Reid, A. J., & Langhorne, J. (2021).
Analysis of *pir* gene expression across the *Plasmodium* life cycle.
Malaria Journal, *20*(1), 445. https://doi.org/10.1186/s12936-021-03979-6

Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).
Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.
When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.
Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

The *Plasmodium interspersed repeats* (*pir*s) are described as the largest multigene family found in the genome non-Laveranian species of the malaria parasite, *Plasmodium*; however, their function remains unknown. In this thesis, I analyse *pir* transcription across the lifecycle of three *Plasmodium species*, using published and unpublished data, to identify potential targets for investigating *pir* function. RNAseq data from the rodent parasite *Plasmodium berghei* demonstrated that *pir* transcription peaks in the liver merosomes/merozoites, the asexual blood stages, and the male gametocytes. Detailed analysis of synchronous *in vivo* asexual blood stages of another rodent parasite, *Plasmodium chabaudi chabaudi,* showed that *pir* transcription peaks in the late trophozoites. Schizont stages of both rodent species demonstrate an enrichment for L-form *pir*s, linking this sub-family to schizogony. With *de novo* transcriptomics methods applied to the *pirs* of the human parasite *Plasmodium vivax*, I observed that transcription of *pir*s in the mosquito stages was minimal, in agreement with the data from *P. berghei*. The *P. vivax* asexual blood cycle pattern of *pir* transcription also showed high expression, but timings contrasted to that of *P. c. chabaudi* as transcription peaked in late schizonts/early rings. Overall, the results suggest that *pirs* are important during the lifecycle stages in the blood, and not in the insect vector. In all three *Plasmodium* species, I found the conserved ancestral *pir* to be highly transcribed, indicative of an essential function. In conclusion, by collating data from a range of sources I have identified consistent patterns of *pir* transcription between *Plasmodium* species at different stages of their lifecycle. Based on these results, I have postulated on the possible role of *pir*s in the biology of the parasite. This study will enable more targeted future studies into elucidating the function of this elusive multigene family.

Acknowledgement

The five years of this PhD have been a lot to handle, so I have a lot of people to thank for the scientific and the emotional support I needed to get here.

Firstly, thank you Jean and George for being my supervisors and taking me on as their student. My project had quite a tumultuous progression and ended up pretty different from how it started out, converting from mostly experimental to entirely computational, and both of you were central to helping me through this and suggesting new ways to do the research that suited my health. Jean, thank you for keeping me focussed on my questions and the aims at task, which my 'bioinformatics brain' often steered me away from. You've also kept me paying attention to my own health, and I still recall fondly when you said to me; "Look after yourself, you're worth it". You also supported me to go to Australia which was a real highlight of my experience and had me seriously considering moving there permanently. George, thank you for contributing so many interesting ideas and discussions even though my visits to Imperial were infrequent!

Thank you to Moritz Treek, Edgar Deu, Adam Reid and Andrea Crisanti for being on my thesis committee. Not every PhD programme has thesis committees, and I am so grateful that mine did, as the advice and support I received was essential for my project's development, and just generally feeling like I had a bunch of great people who had my back.

Thank you Deirdre Cunningham and Adam Reid, for taking on close mentorship roles for me, especially later in my project. I really needed the support that you both provided and I do not think I could have produced a thesis remotely to the quality I have without it. My first ever publication was co-authored with you Deirdre, with Adam as a senior author, and that holds a special place in my heart. Indeed, the paper is framed on my wall, if you needed any proof!

My time in the lab and on the computer were assisted by many additional colleagues and friends who mentored and guided me through difficult problems, techniques, and times. So, thank you Chris, John, Jade, Claudia, Sarah Kelly, Sarah Manni, Bob, George Young, Prisca, Caroline, Dina, and Katrien. My fellow PhD students have been there to

understand the very particular student problems, so thank you Irene, Maria Giorgalli, Sara Amis, Kelvin, and Alex.

I've also worked with many people who were doing stuff a bit too different to my project for us to work directly together, but who were valued friends during my time at the Crick and Imperial labs, including (but not limited to) Jenny, Morgane, Faith, Matt, Ellen, Lauren, Tib, Carlos, Lucy, Yaw, and Melina – Thank you! The datasets I used were produced by an effort of many, but spearheaded by Audrey Vandomme and Thibaut Brugat, so thank you both for contributing to this thesis in a way I'm sure neither if you expected to. Thank you to the other parasite labs in the building for the Api Hours, and the kind words everyone had for my presentation about bird and lizard malarias! My previous mentors have also led to me to this moment, so thank you Gavin Wright, Pete Bull, Brian Hendrich and John Macauley. From the non-academic side, thank you to the Crick student team and in particular Caroline Ransom for help with deferring and other student-y pains I've experienced.

My time at the Crick has enriched my life with great friends who I hope will remain close to me long after I have left the exulted halls, and they deserve a huge shout out for everything their love, kindness, and patient listening to my whinging, has done to help me get by. Not to forget, of course, the much-needed coffee(/drink) breaks and squash games! So, thank you Fernanda, Louise, Tom, Gen, Melissa, Margaux, Theo, Cristina, Jorge, Clovis, Kevin, Simran, Justina, Lisa, Windie, Tereza, Emma, and many other friends.

Beyond the world of The Francis Crick Institute, I have had many curiously looking at the weird pressures put on me, but who have been there to support me without a second thought. First-and-foremost, Abby, thank you for supporting me perhaps more than anybody else has! Listing my friends outside of the Crick would be an impossible task, so to be very succinct a thank you is deserved for Pol, Tomek, Chris Cundy, Eleni, Eddie, Hannah, Robbie, David, Jack Jack, Tara, Etienne, Jasmine, Famelton, Evan, Alasdair, Jodie, Gerard, and Andy (even though you once left without saying goodbye).

My family too have had my back no matter what, even if they do not always understand what's going on in the mysterious academic world. Thank you to all of them, but especially to Mum and Dad who embody how you never truly retire from being a parent.

For visiting me and sending care packages, for looking after me in Ireland and always being there for a phone call, for never doubting me and literally being my alarm clock as if I'm still at school, thank you so much. I love you.

Two special babies who have even less of an understanding of what is going on are my precious furballs; thank you old baby Odin and naughty chonk Jupiter. I'm sorry that this PhD has meant I've needed to ignore or leave you alone for hours, I hope it was worth it.

I did it, everyone!

Table of Contents

Abstract **4**
Acknowledgement **5**
Table of Contents **8**
List of figures **10**
List of tables **13**
Abbreviations **14**
Chapter 1.Introduction **16**
1.1 Malaria parasites and disease **16**
1.2 The Apicomplexan parasite responsible for malaria **17**
1.2.1 The human, rodent and simian malarias 17
1.3 The life cycle of *Plasmodium sp.* **18**
1.3.1 Malaria parasites in the mosquito – from infected blood meal to infective insects 18
1.3.2 Exo-erythrocytic stages of malaria in the host 19
1.3.3 The asexual and sexual stages of parasites in the blood 19
1.3.4 The importance of understanding vector stages of development 20
1.4 Multigene families in biology and malaria **21**
1.4.1 The evolution of multigene families 21
1.4.2 Multigene families and superfamilies across species 22
1.4.3 The importance of multigene families for malaria parasites – the *var* genes 22
1.4.4 "Hedge-betting" and immune evasion 23
1.5 The *Plasmodium interspersed repeats* multigene family **24**
1.5.1 The sub-families of the *pir*s 25
1.5.2 The evasiveness of understanding *pir* function 27
1.6 High-throughput RNA sequencing has shed light on *pir* transcription across the life cycle **29**
1.6.1 *P. c. chabaudi* AS transcriptome analysis suggests a role of *pir*s in pathogenicity 29
1.6.2 RNAseq data from other species demonstrates patterns among *pir* transcription 31
1.7 The aims and hypothesis of this thesis **32**
Chapter 2.Materials and Methods **34**
2.1 *P. c. chabaudi* AS culturing and RNAseq extraction **34**
2.1.1 Attribution of labour 34
2.1.2 Mice 34
2.1.3 Mosquitoes 34
2.1.4 Parasite maintenance 35
2.1.5 RNA extraction and sequencing 35
2.2 RNAseq analysis of the *P. c. chabaudi* AS and *P. berghei* data **36**
2.2.1 Dataset download and processing 36
2.2.2 Data filtering and analysis in R 37
2.2.3 Statistical analysis of differential expression in R 37
2.2.4 Deconvolution of *P. c. chabaudi* AS bulk RNAseq 38
2.3 Phylogenetic analysis of the *pir* 'ancestral' sequences **38**
2.3.1 Transcriptional comparisons of the ancestral *pir* across the Plasmodia 39

2.4 *De novo* transcriptomic analysis of *P. vivax* RNAseq data Methods . 39
2.4.1 *P. vivax* data download and adapter trimming 39
2.4.2 *P. c. chabaudi* AS RNAseq pre-processing, *de novo* transcriptome assembly and analysis 40
2.4.3 *P. vivax* RNAseq pre-processing and *de novo* transcriptome assembly40
2.4.4 *P. vivax pir* transcript detection 41
2.4.5 Creating networks of PIRs and sub-family assignment..... 42
2.5 Other packages used in R throughout analysis 43
2.6 Availability of R code 43
Chapter 3.Results - Analysis of *P. berghei pir* transcription across the life cycle 44
3.1 Introduction..... 44
3.2 Aims 46
3.3 Results..... 47
3.3.1 RNAseq samples from different stages of the *P. berghei* life cycle robustly cluster by the parasite developmental stage 47
3.3.2 The total transcription of *pir*s across the life cycle of *P. berghei* consistently oscillates with given sub-family enrichment at certain stages 56
3.3.3 The ancestral *pir* gene is consistently transcribed at high levels across the *Plasmodium* genus. 64
3.4 Conclusions 67
Chapter 4.Results - Transcription of the *pir* genes across the intraerythrocytic development cycle of *P. c. chabaudi* AS..... 73
4.1 Introduction..... 73
4.2 Aims 77
4.3 Results..... 78
4.3.1 RMT samples are mildly asynchronous but still dominated by expected stages of the IDC..... 78
4.3.2 Transcription of *pir* genes peaks during the later trophozoite stages.. 83
4.3.3 Differential transcription patterns of the pir sub-families and the relationship to chronic and acute-associated loci 87
4.3.4 Common patterns of pir transcription between *P. c. chabaudi* AS and *P. berghei* 95
4.4 Conclusions 101
Chapter 5.Results - Analysing the *P. vivax pir* transcriptome using *de novo* transcriptomics..... 110
5.1 Introduction..... 110
5.2 Aims 113
5.3 Results..... 114
5.3.1 *De novo* transcriptomes of *P. c. chabaudi* AS data recover known *pir*s114
5.3.2 Asexual blood cycle RNAseq of *P. vivax* can produce many *de novo* assembled *pir* transcripts 121
5.3.3 *De novo pir* transcripts both cluster within previously described sub-families and describe novel groups 130
5.3.4 Transcription of *de novo pir*s is both spread over the major sub-families and diverse between parasite isolates 137
5.4 Conclusions 142
Chapter 6.Conclusion 151
Chapter 7.Reference List..... 162
Chapter 8.Appendix 197

List of figures

Figure 1.1 – Phylogeny and *pir* sub-families of the *Plasmodium* genus.....26
Figure 3.1 - Schematic of the life cycle stages of *P. berghei* covered by the different RNAseq publications used in this investigation.....46
Figure 3.2 - Hierarchical clustering of the P. berghei RNAseq samples in TPM from E03 and E08.....49
Figure 3.3 - Density plots of the RNAseq sample ln(TPM+1) values for E06 and E09. ..50
Figure 3.4 – Principal component analysis (PCA) plot of the genome TPM data coloured by the different stages and shapes determined by the originating experiment. 51
Figure 3.5 - Biplot of the *P. berghei* experimental samples, highlighting the top 15 genes influencing the PCA results. 53
Figure 3.6 – Pearson correlation plot (-0.04-1.00) of each transcriptome calculated from genome TPM values for each stage from each experiment.....54
Figure 3.7 - Transcription of gametocyte markers among the *P. berghei* experiments and life cycle stages. 55
Figure 3.8 - The median TPM of *pir* genes (right) and 134 randomly selected non-*pir* genes (left) across the life cycle. 57
Figure 3.9 - Transcription of *pir* genes throughout the *P. berghei* life cycle, split by sub-family designation.....58
Figure 3.10 - Proportions of total *pir* TPM contributed by different sub-families across the life cycle.....59
Figure 3.11 - Heatmap of the log-fold change of the *pir*s and *pir* sub-families between sequential and selected stages of the *P. berghei* life cycle.....61
Figure 3.12 - Heatmap of the log fold changes of individual *pir* genes with statistical evidence for differential expression (FDR < 0.05) between given stages of the *P. berghei* life cycle.....62
Figure 3.13 - Alignment of the ancestral *pir* from multiple *Plasmodium spp.* shows high degrees of conservation. 65
Figure 3.14 - Phylogenetic and gene expression analysis of the ancestral pir gene. 66
Figure 4.1 – Schematic of the RNAseq extraction across the asexual life cycle for both SBP and RMT infections of *P. c. chabaudi* AS.....76
Figure 4.2 – Proportions of parasite blood stages for each timepoint. 79
Figure 4.3 – Variation within and between the RMT/SBP asexual blood cycle samples.80

Figure 4.4. Phaseograms of maximum/minimum transcription of every gene of the transcriptome across the twenty-four-hour asexual life cycle of *P. c. chabaudi*. 82
Figure 4.5. PCA plot of the RNAseq samples using the *pir* gene TPM values. 84
Figure 4.6 – Bar chart of *pir* gene TPM across the asexual cycle and RMT schizont cultures of *P. c. chabaudi* AS transmitted by RMT or SBP. 85
Figure 4.7 – Bar chart of TPM of the ancestral *pir* gene *PCHAS_0101200* across the asexual blood cycle and RMT schizont cultures. 86
Figure 4.8 – Bar chart of TPM of the SBP-high *pir* gene *PCHAS_1100300* (an individual *pir* massively upregulated in SBP samples compared to MT in Spence et al, 2013) across the asexual blood cycle and RMT schizont cultures. 87
Figure 4.9 – Phaseograms of the z-scores of the *pir* genes compared between RMT and SBP. 89
Figure 4.10 – Heatmap of the log(TPM+1) values of the *pir* genes across the twenty-four-hour cycle of *P. c. chabaudi* AS in both RMT and SBP transmitted infections, and RMT schizont cultures. 90
Figure 4.11 - Heatmap of the log(TPM+1) values of the ChAPL/AAPL *pir* genes across the twenty-four-hour cycle of *P.c. chabaudi* AS in both RMT and SBP transmitted infections, and RMT schizont cultures. 92
Figure 4.12 - Heatmap of the log fold change of gene set transcription comparing consecutive time-points of the asexual cycle for RMT and SBP infections. 94
Figure 4.13 – Heatmap of the log fold change of gene set transcription comparing the same time points of RMT and SBP infections. 96
Figure 4.14 – Schematic of the approximate changes of *pir* gene transcription across the IDC for RMT (grey) and SBP (orange) parasite populations. 97
Figure 4.15 – Filled bar chart showing the proportion of *P. c. chabaudi* AS non-ancestral *pir* sub-family transcription across every sample. 98
Figure 4.16 – Filled bar chart showing the proportion of *P. berghei* non-ancestral *pir* sub-family transcription across the non-mixed asexual samples of E01 and E03 (see Table 3.1); 99
Figure 5.1 – Flow chart of the analysis of *P. vivax* RNAseq datasets to produce and investigate *de novo* transcriptomes of the *pir* genes. 119
Figure 5.2 – The numbers of known *P .c chabaudi* AS *pir*s covered by *de novo* transcripts from the assemblies of a 14h rMT RNAseq sample. 120
Figure 5.3 – Percentage of BUSCOs present in the assemblies. 124

Figure 5.4 – The number of *de novo pir* transcripts with predicted *pir* ORFs at least 200aa long against the total number of transcripts for different *de novo* assemblies of *P. vivax* RNAseq. 125
Figure 5.5 – Percentage of the *de novo* transcriptome transcribed over 1 TPM against the total size of the transcriptome. 126
Figure 5.6 - Visualisation of the 30% identity and 25aa length similarity PIR network, showing the top twenty clusters, using an MCL inflation value of 1.4. 128
Figure 5.7 - Visualisation of the 30% identity and 25aa length similarity PIR network, showing the assigned Sal1 sub-family sequences. 129
Figure 5.8 – The distribution of the Sal1 sequences across the largest twenty-five clusters of the 30% identity and 25aa length similarity PIR network, clustered using an MCL inflation value of 1.4. 131
Figure 5.9 – The composition of the largest ten clusters among the experimental and known *de novo* PIRs sequences from the 30% identity and 25aa length similarity PIR network, clustered using an MCL inflation value of 1.4. 133
Figure 5.10 – Transcription of *de novo* assembled *pir* genes from RNAseq alignment to the assemblies for E07 and E09. 135
Figure 5.11 – Transcription of *de novo* assembled *pir* genes from RNAseq alignment to the assemblies for E06. 138
Figure 5.12 – Transcription of *de novo* assembled *pir* genes from RNAseq alignment to the assemblies for E01 across the IDC, including patient isolate 1 (A) and patient isolate 2 (B). 140
Figure 6.1 – Summary figure of *pir* transcriptional patterns observed in this thesis across the life cycle of *P. berghei*, *P. c. chabaudi* AS, and *P. vivax*. 153

List of tables

Table 3.1 - Published studies from which P. berghei RNAseq datasets were used in this chapter, including the stages of P. berghei extracted and the experimental code used to identify sample origins. 48
Table 3.2 - Published studies from which transcriptome data was taken from different *Plasmodium spp.* and used in Figure 3.14. 67
Table 5.1 – *De novo* transcriptome assembly statistics for the *pir* transcripts from the rMT and SBP data combined, using different assembly programs. 115
Table 5.2 - *De novo* transcriptome assembly statistics for the *P. c. chabaudi* AS *pir* transcripts from a 14h rMT sample, using different assembly programs and CD-hit-est redundancy removal thresholds (90%, 95% and 99%). 118
Table 5.3 – Summary of the publications from which *P. vivax* RNAseq data was taken and used for *de novo* assembly. 122
Table 5.4 – The largest twenty clusters from the 30% identity and 25aa length similarity PIR network, clustered using an MCL inflation value of 1.4. 132
Table 8.1 – RNAseq read files of *P. vivax* and which were combined for *de novo* transcriptome assembly. 210

Abbreviations

AAPL	Acute-associated *pir* loci
BCR/TCR	B-cell/T-cell Receptor
BUSCO	Benchmarking Universal Single-Copy Orthologue
ChAPL	Chronic-associated *pir* loci
CQ	Chloroquine
ETRAMP	Early transcribed membrane protein
FDR	False-discovery rate
FPKM/RPKM	Fragments/Reads per kilobase of exon per million reads mapped
HMM	Hidden Markov model
IDC	Intra-erythrocytic development cycle
Ig	Immunoglobulin
iRBC	Infected Red Blood Cell
mapk2	*Mitogen-activated kinase 2*
mss	*male secreted short*
Nek4	*NIMA-related kinase 4*
PCA	Principal component analysis
PfEMP1	*Plasmodium falciparum* Erythrocyte Membrane Protein 1
pir	*Plasmodium interspersed repeats*
PVM	Parasitophorous vacuolar membrane
RMT	Recently Mosquito Transmitted
S/L-form	Short/Long-form
SBP	Serial blood passage

scRNAseq	Single-cell RNAseq
Sdic	*Sperm-specific dynein intermediate chain*
SRA	Sequence Read Archive
TM	Trans-membrane
TPM	Transcripts per million
vir	[*Plasmodium*] *vivax interspersed repeats*

Chapter 1. Introduction

1.1 Malaria parasites and disease

The *Plasmodium sp.* represent some of humanity's most formidable infectious foes from across history and into the present day. The protozoan parasites have afflicted humanity for so long that modern human genomes have been shaped by their influence (Allison, 1954; Haldane, 1949). The UN states that malaria obstructs many of the Millennium Development Goals (Hunt, 2007). While mortality from malaria has been reduced significantly over the last few decades, efforts have been unable to maintain this decrease in recent years. Because of the plateauing effectiveness of interventions and the disruptive COVID-19 pandemic, in 2020 there were 14 million more malaria cases and 69,000 more deaths than the year prior (WHO, 2021). In total this means that there were 241 million cases of malaria infections in 2020, culminating in 627,000 deaths. The majority of these are due to *Plasmodium falciparum*, the deadliest, but by no means the only, human-infecting malaria parasite. Hence it is premature to state that malaria is under control, especially in the face of poorly predictable threats such as global warming and antimalarial resistance.

The burden of malaria is not evenly distributed across the world, or within populations. 85% of infections in 2020 developed in Africa, and two African nations bore over 10% of the world’s malaria cases each - Nigeria and Democratic Republic of the Congo. In fact, Nigerians alone experienced 27% of the global number of infections in 2020. Other worldwide foci of malaria infections include South-East Asia, the Western Pacific and Central and South America. Again, within these regions, there were countries that bore the brunt of the malaria caseload (India, Papua New Guinea and Brazil/Venezuela/Colombia respectively) (WHO, 2021). Just a glance at these statistics reveals that malaria remains a disease of poverty (Tusting et al., 2013). Although local elimination is both possible and achievable, it is still only a long-term goal for many poorer nations. Even when a region does eliminate malaria, as Europe, Sri Lanka, Argentina, China, and many others, have done, this status is delicate and must be rigorously maintained. Notably Sri Lanka was on the precipice of eliminating malaria within its borders back in 1963 but lost the political will and the disease resurged. Decades later, in 2016, Sri Lanka was declared malaria free and the country has kept this status ever since (WHO, 2016).

The potentially deadly symptoms of malaria vary greatly both between and within different *Plasmodium sp.* strains, and furthermore differ from host to host. This can be seen even with *P. falciparum*, which can cause asymptomatic infections, mild fevers, coma, or lethal respiratory distress. Multiple factors have been identified to affect the symptoms of an infection in different models, including host immunity (Langhorne et al., 2008), host nutritional state (Mancio-Silva et al., 2017) and parasite genetics (Rorick et al., 2013), however much variation in virulence remains unexplained. A consistent pattern is that the least immune are the most susceptible, and in endemic regions this is the burden of the youngest in society. Children under the age of five accounted for 77% of worldwide deaths in 2020 (WHO, 2021).

1.2 The Apicomplexan parasite responsible for malaria

Plasmodium parasites are an ancient lineage, and hence they are highly diverse in both their pathology and the animals they can infect. The earliest evidence of *Plasmodium* comes from amber-preserved remains of a greater than 15-million-year-old mosquito infected with structures that resemble oocysts and sporozoites (Poinar, 2005). The Plasmodia are Apicomplexan parasites, part of a huge phylum that includes *Toxoplasma spp., Cryptosporidium spp. Babesia spp.*, and *Eimeria spp* (Levine, 1970). Drilling deeper into the taxonomy, the Order Haemosporida encompasses the many different parasites considered historically to cause malarial disease. The detailed phylogenies are often changing with new studies and are by no means certain. Modern evidence suggests that the *Plasmodium* genus itself is polyphyletic; split across different clades, with other genera more closely related to some species than other *Plasmodium spp.* (Galen et al., 2018). This nomenclature disaster helps explain the diversity of malaria parasites; many of them are not at all closely related to each other. Among the Haemosporidians less closely related to human malarias are parasites that infect birds, lizards and deer.

1.2.1 The human, rodent and simian malarias

As we zero in on the human malarias, we see that they too are split into different clades: the Laverania sub-clade that includes *P. falciparum* and other closely-related simian

malaria parasites, and another clade that includes the human malaria parasites *P. vivax*, *P. knowlesi, P. ovale,* and *P. malariae*, simian malaria parasites, and small mammal-infecting malaria parasites (Galen et al., 2018). So, as well as being the most dangerous, *P. falciparum* is phylogenetically distinct from the other human malaria parasites.

The clade of parasites that infect small mammals includes some of the most important models of malarial disease. Extracted from West African thicket rats, the rodent malarias *P. berghei*, *P. chabaudi* and *P. yoelii* are able to infect laboratory animals and thus provide a tractable model for the study of malaria parasites (Landau, 1965; Landau & Killick-Kendrick, 1966; Vincke & Lips, 1948; Yoeli & Hargreaves, 1974). With the most up-to-date phylogenies in mind, these models are most closely related to non-*falciparum* human malaria parasites, including *P. vivax*.

1.3 The life cycle of *Plasmodium sp.*

The success of *Plasmodium sp.* as parasites is closely linked to their life cycle. They are vector-borne protozoa, transmitted between hosts by flying, biting insects (Ross, 1898). All human-infecting Plasmodia are transmitted by mosquitoes, a creature whose relationship with malaria is central to understanding the parasite and the disease, as well as how to control it.

1.3.1 Malaria parasites in the mosquito – from infected blood meal to infective insects

When a mosquito feeds on infected blood, it may take up the *Plasmodium* sexual stages (male and female gametocytes) which differentiate into gametes. The male gametocyte undergoes exflagellation upon exposure to the internal conditions of the mosquito (such as a lowering of the temperature and chemical factors), producing multiple flagellated, motile male gametes (Billker et al., 1998; Kligler & Mer, 1937; Laveran, 1881; Maccallum, 1897). A male and female gamete fuse in the mosquito midgut, within the bloodmeal mid-digestion, forming a zygote, which further differentiates into the ookinete (Grassi, 1901; Maccallum, 1897). The ookinete is a highly specialised motile form of the parasite, with the purpose of breaking through the mosquito midgut barriers to penetrate the epithelium (Sieber et al., 1991). If the ookinete is successful then it will implant itself

and become the oocyst, which expands like a balloon of proliferating parasite. Inside the oocyst many individual sporozoites develop – the stage of the life cycle which is infective to humans (and other *Plasmodium sp.* hosts) (Grassi, 1901; Ross, 1897). The oocyst bursts after several days and the sporozoites migrate through the mosquito haemocoel to the salivary glands, where they reside until the vector takes a new blood meal. Upon biting a host, the mosquito secretes saliva laden with sporozoites.

1.3.2 Exo-erythrocytic stages of malaria in the host

The malaria parasite has now entered the host, and must move from the skin/circulation, where it was injected, to new tissue for further development (Boyd & Kitchen, 1939). Across other Haemosporidian parasites, these exo-erythrocytic stages can form in endothelial cells and monocytes (James & Tate, 1938). However, in the known clades that infect humans, mice, and primates, sporozoites invade hepatocytes and begin to grow, forming exo-erythrocytic schizonts (Shortt et al., 1948; Shortt & Garnham, 1948). This time the proliferating developmental stages inside the ‘mother’ parasitic cell are merozoites – the red blood cell infective stage. These cells are released through the budding off portions of membrane, termed merosomes (Sturm et al., 2006). When these burst forth from the exo-erythrocytic schizont they are released into the sinusoidal vessels and can invade erythrocytes that they encounter – forming infected red blood cells (iRBCs).

1.3.3 The asexual and sexual stages of parasites in the blood

The liver-borne merozoite instigates the asexual blood cycle of development for the parasite, in which the invader matures from a ‘ring’ stage, into a trophozoite, and finally another schizont stage. Thus, the infected cells support a progression of clonal proliferation of new merozoites which can perpetuate the infection upon release. Some of the iRBC, at an unknown stage of development, will sexually commit to become gametocytes. This then completes the cycle, as these gametocytes can reproduce in a feeding mosquito.

The asexual blood cycle is the cause of clinical symptoms, and the association between the timing of the cycle and paroxysm of symptoms in patients was observed early in the

history of malariology (Golgi, 1886). Furthermore, there is evidence that some of the deadliest symptoms of some malarias are a consequence of infected red blood cell sequestration around the circulation of the body, such as the occurrence of *P. falciparum* parasites stuck to the brain endothelium in coma patients (Marchiafava & Bignami, 1892). This is believed to be an adaptation to avoid clearance in the spleen, which will selectively target the iRBC that try to pass through (Barnwell et al., 1983). Parasite sequestration and the proteins underlying it are discussed in 1.4.3.

1.3.4 The importance of understanding vector stages of development

Although the asexual blood cycle receives a lot of study because of its relevance to clinical malaria, investigating the remaining parasite developmental stages is essential for controlling other aspects of *Plasmodium sp.* infection and transmission. The growth of *Plasmodium* in the mosquito is when they are at their most vulnerable, as the population undergoes multiple bottlenecks until eventually proliferating again in the oocyst. The most prominent bottlenecks of transmission occur during parasite fertilisation and attack by the mosquito immune response. The success rate of fertilisation between *Plasmodium sp.* gametes in the mosquito midgut is low, and almost immediately after entering the vector the parasite cell population drops several-fold (Poudel et al., 2008). Even if the zygote is successfully formed and can develop into an ookinete, the mosquito possesses several immunological barriers to the parasites as they attempt to migrate out of the midgut. Often infections of several thousand cells are reduced to a single cell by this stage or cleared altogether. These present the greatest population bottlenecks of the entire malarial life-cycle; gametocyte-to-ookinete (greatest fold change) and ookinete-to-late-oocyst (fewest surviving cells) (Vaughan, 2007). The exo-erythrocytic stages of the parasite, which initiate the asexual blood cycle, can be central to the chronicity of parasite infections. Most notably the human malaria *P. vivax*, and some other species, can develop into a long-term quiescent liver form known as the hypnozoite. These cells can lie dormant for months or years until they reactivate and initiate a recrudescent infection (Krotoski et al., 1980, 1982).

1.4 Multigene families in biology and malaria

The molecular toolkits employed by pathogens such as *Plasmodium spp.* are ultimately determined by their genes. Scrutinising DNA sequence reveals that many genes cluster together, demonstrating that genes in a genome do not develop independently of one another. Groups of related genes form gene families, and usually when they are several members or greater in size (there is no set definition) they are termed multigene families.

1.4.1 The evolution of multigene families

How different loci gain related sequence has been debated often over the years, and the mechanism likely depends on the genes in question (Nei & Rooney, 2005). Ultimately the fundamental unit of multigene generation is gene duplication from an ancestral gene, creating a nascent gene family of two members. Further duplication expands the family, and the same gene family can range hugely in member number between different species. Until the early 1990s, the prevailing thought was that multigene families subsequently developed through 'concerted evolution' – in which duplications were homogenised by constant swapping of sequence via partial gene conversion (essentially the replacement of homologous sequence from one locus to another). This explained the finding of high intraspecies and lower interspecies conservation of rRNA genes at different loci, first described in *Xenopus sp*. (Brown et al., 1972). However, this did not explain multigene families in which there was high polymorphism between members, and where sub-lineages persisted between different species (i.e. some groups of genes which have higher interspecies than intraspecies conservation within the same multigene family). A new model was proposed, the 'birth-and-death' model, in which genes duplicate and then diverge in sequence, with gene conversion occurring more rarely. As the family grows the members further duplicate and diverge with evolutionary time, leading to highly similar but distinct genes dotted around the genome. Some of these genes lose function and become pseudogenes – strongly predicted to be defective due to premature stop codons or frameshifts (Nei et al., 1997). With this paradigm, one can go beyond species phylogenies and analyse the phylogenies of genes themselves. Multigene families derived from birth-and-death processes will have a phylogenetic tree that links them to each other by evolution, and by tracing the tree right to the root you can see that this genealogy is derived from a predicted common ancestral gene.

Sometimes even a highly basal, but extant, gene can be found, which may be the surviving form of the original gene. With this model, the sub-lineages are understood to have been formed from the individual lineages of genes derived from the ancestral sequence, and they can be classified into distinct units.

1.4.2 Multigene families and superfamilies across species

A number of similar multigene families can be grouped into larger superfamilies. Many are reservoirs of genetic diversity since they harbour strong redundancy of function – so if an individual gene loses/gains function the operation of the organism is minimally impacted. The immunoglobulin (Ig) superfamily is one example; the most populous gene family of the human genome with 765 known members (Lander et al., 2001). All these genes share a common structural fold and most have a role in cell signalling/receptor binding in immunity. However, their function can escape from this paradigm, such as IZUMO1 which is involved in cell-to-cell communication during mammalian zygote fertilisation (Inoue et al., 2005). This superfamily is ancient among the Domains of life, having been found in *Streptococcus agalactiae* and other bacteria (Bateman et al., 1996).

1.4.3 The importance of multigene families for malaria parasites – the *var* genes

It is striking that *Plasmodium spp.* also have expansive gene families, particularly as their genomes only contain approximately 6000 genes in total. In fact, the overall lack of redundancy is demonstrated in knock-out screens of *P. berghei*, in which around 45% of the ‘core’ genome are found to be essential for blood-stage development (Bushell et al., 2017). However, *Plasmodium* indeed utilises numerous gene families, some of which are essential for infection. The best studied example is the *var* gene family of *P. falciparum*, which encode the *Plasmodium falciparum* Erythrocyte Membrane Proteins 1 (PfEMP1). These genes are important for sequestration of the parasite in the host; the parasite will express and export PfEMP1 to the infected cell surface which will attach to ligands in different endothelial tissue (Su et al., 1995). Since they are on the surface of the iRBC they are exposed to the immune system and are targets of attack. This is where the gene family nature of the *var*s becomes important. As expression of a given PfEMP1 protein on the iRBC leads to their elimination by adaptive immunity, the

population can ‘switch’ to a new PfEMP1 as cells expressing this new protein are selected for (Smith et al., 1995). Only one *var* gene will be transcribed in an iRBC at a time, enabling this behaviour. The structure of the *var*s is modular, comprising distinct sequence blocks, and since the *var*s are positioned in sub-telomeric regions, these blocks of sequence can be recombined between each other, permitting great degrees of diversity of the genetic code (Smith et al., 2000). Although the underlying sequence can change, the function of the PfEMP1 (that has been selected for) remains cell-to-cell binding, showing a structural conservation where sequence is divergent (Lau et al., 2015). However different *var* genes can target different ligands. This is observed most starkly in malaria of pregnancy, characterised by the accumulation of stationary parasites in the placenta (Clark, 1915). In these infections *P. falciparum* expresses a particular conserved *var* locus (*var2csa*) which binds CSA protein in the placenta (Fried & Duffy, 1996). As with the immunoglobulins, the PfEMP1s do not all function exclusively in immune interactions. One such protein is produced in the sporozoite to enable liver cell invasion; a member of the *var*s which must have a divergent function from the rest but retains the cell-to-cell binding activity (Zanghì et al., 2018).

1.4.4 “Hedge-betting” and immune evasion

Multigene families such as the *var* can enable ‘hedge-betting’ strategies (i.e. investing in a plethora of options instead of relying on one) in organisms when entering a new environment or situation. By expressing as many different variants of a protein as possible, the hope is that at least some will be fit and enable the survival of individuals or a population. The adaptive immune system provides another example here, as hedge-betting applies to the purposefully generated extreme diversity of the B-cell and T-cell receptors (BCR/TCR respectively). By producing numerous different sequences, and selecting for those that are still functional, the immune system has covered as many bases as possible for binding a pathogen when it attacks the body. Working in opposition to this process is the antigenic diversity of pathogens, such as that encoded by the *var* genes. It is hypothesised that when the pathogen enters a new human host it has no idea which PfEMP1 they have already developed immunity to, if any. If the population expresses many different variants, then there is more likely to be a sub-population that survives and can proliferate. Indeed, NF54 strains grown *in vitro* show upregulation of only one *var* gene (*var2csa*), but populations of NF54 from controlled human infections, via sporozoite injection, demonstrate wide ranging *var* transcription

across sub-groups in comparison (Bachmann et al., 2016). This process then repeats when the major PfEMP1 becomes the target of antibody, and a new surface protein is fit. This strategy can be severely hampered by interfering with the monoallelic expression of similar hedge-betting proteins, as shown with the variant antigens of other parasites *Giardia lamblia* and *Trypanosoma brucei*. When the pathogen cells produce multiple different proteins at once then the host can quickly target them all (Aresta-Branco et al., 2019; Rivero et al., 2010).

1.5 The *Plasmodium interspersed repeats* multigene family

While the Laverania sub-genus possesses the *var* genes, another multigene family is found in the other, neighbouring clade of *Plasmodium spp*. Across the sequenced genomes of the parasites of this clade, a multigene family known as the *Plasmodium interspersed repeats* (*pir*s) has been detected. Members of the family generally share a three-exon structure; a short first exon, a long second exon with a central variable region and trans-membrane domain (TM), and alpha helices conserved both in sequence and position. The gene family was initially discovered as clusters of closely related genes in the sub-telomeric sequence of *P. vivax* (termed the *vivax* interspersed repeats - *vir*s) and hypothesised to be immune interactors in species lacking *var* genes. Since the *vir*s were described, their sequence was recruited to predict fellow members across the genomes of different malaria species, using hidden Markov models (HMMs) (Janssen et al., 2004). The gene family has been found in the genomes of rodent, primate, and human-infective malarias at highly variable copy numbers (del Portillo et al., 2001; Janssen et al., 2002; Otto et al., 2014; Pain et al., 2008). For some species the number of *pir*s is relatively low, such as the 134 of *P. berghei*, while for others it can be in the thousands, such as *P. ovale, P. yoelii* and *P. vivax* P01 strain (Auburn et al., 2016; Otto et al., 2014; Rutledge et al., 2017). For these latter pathogens the *pir* family astonishingly makes up about one-sixth of the total number of genes. Additionally, four similar genes have been discovered in avian malarias (Böhme et al., 2018). Hence, *pir*s are described as the most widespread gene family in the genomes of malaria parasites (Janssen et al., 2004). Their presence in the Laverania has been a matter of dispute, due to having both similarity and major differences with the *rifin/stevor* multigene family (Frech & Chen, 2013; Janssen et al., 2004). The crystal structure of a *P. c. chabaudi AS*

PIR fragment further distanced the two families, as it showed no discernible similarity to RIFIN or STEVOR structure of *P. falciparum* (Harrison et al., 2020).

1.5.1 The sub-families of the *pir*s

Constructing phylogenies of the *pir*s across species has shown that they can be divided into sub-families; gene groups which likely share a more recent common ancestor than the overall common ancestor. In the rodent malarias, the *pir* phylogeny separates into two groups with distinctively different average gene lengths (see Figure 1.1). The longer *pir*s, of around 1000-2400 aa in length, were termed the Long-form *pir*s, and could be further sub-divided into clades of L1-4 sequences. Their extended sequence was found to be caused by the presence of repeat regions. The Short-form *pir*s encompass many more sequences ranging from 800-1000 aa long predicted proteins and can also be sub-divided into the S1-8 groups. Most of these clades include a few *pir*s from each of the rodent malarias analysed, however they tend to be dominated by representatives from only a few of the species, especially *P. vinckei spp*., *P. c. chabaudi* AS and *P. yoelii* (which all had more annotated *pir*s than *P. berghei*) (Otto et al., 2014; Ramaprasad et al., 2021). This structure holds across the rodent malarias, suggesting that genetic divergence occurred in the common ancestor of these species. Separate sets of sub-families exist in the simian-lineage, designated by letters of the alphabet (see Figure 1.1) (Cunningham et al., 2010; del Portillo et al., 2001). The membership of these families, and even membership of the *pir* family at all, has undergone revision based on publication of a full genome sequence of *P. vivax* and domain structures of the predicted genes (Carlton et al., 2008; Lopez et al., 2013). These *P. vivax*-defined sub-families include *pir* genes from *P. ovale spp.*, *P. malariae* and *P. knowlesi*, with each of these species potentially containing species-unique clades too (Cunningham et al., 2010; Lopez et al., 2013; Rutledge et al., 2017). The updating and refinement of multigene families and sub-families underscore how their definitions can be fuzzy, and their classifications fluid.

The existence of the sub-families and the polymorphism across the *pir*s suggest that *pir* diversity was created through the birth-and-death process of multigene generation (as discussed in section 1.4.1). Since the rodent and simian clades possess separate sub-families of *pir* genes, there is the curious possibility that each lineage had expansions of

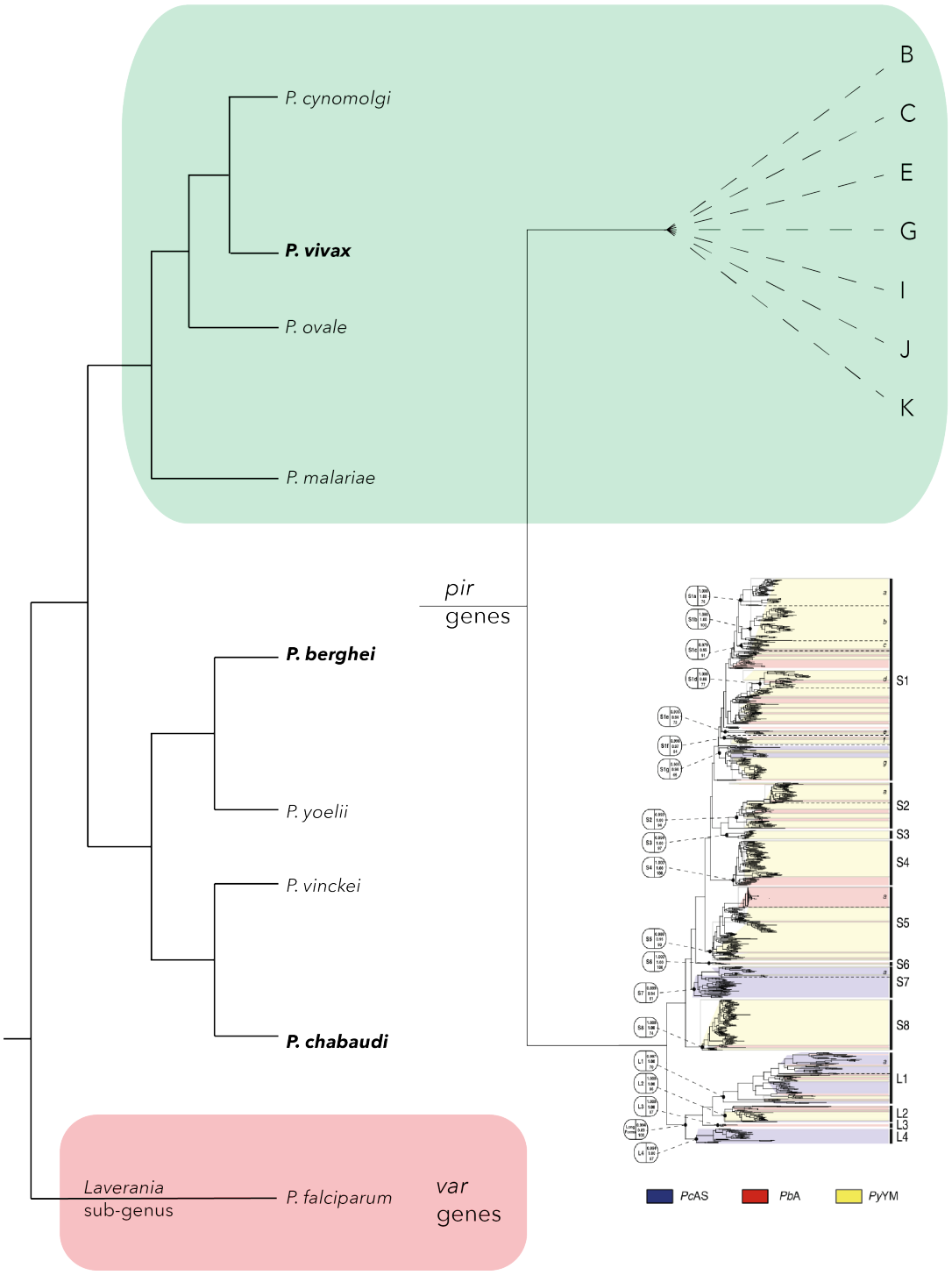


Figure 1.1 – Phylogeny and *pir* sub-families of the *Plasmodium* genus.
The species phylogeny is presented on the left, based on the results of (Galen et al., 2018)**. On the right is the *pir* phylogeny. Highlighted in green are the *P. vivax* lineage species and *pir* sub-families, with dotted lines to indicate a currently undetermined phylogeny. Without highlights are the rodent malaria lineage phylogenies, with the *pir* phylogeny adapted from Otto et al., 2014. S1-8 indicates the Short-form pirs and L1-4 indicates the Long-form pirs. Colours show the *pirs* derived from the genome**

assemblies of each rodent species used, with blue corresponding to *P. c. chabaudi* AS (PcAS), red to *P. berghei* ANKA (PbA), and yellow to *P. yoelii* YM (PyYM). In red highlight is the Laverania sub-genus, an outgroup to this phylogeny, which does not have canonical *pir* genes, but does have multigene families such as the *vars*. With the exception of the Otto et al., 2014, diagram all branch lengths are for illustrative purposes only.

the *pir* repertoire independently of one another, both following a separate trajectory of expansive *pir* duplication from the small copy number of an ancestor.

1.5.2 The evasiveness of understanding *pir* function

The first hypotheses of function were that these genes were part of the antigenic evasion mechanism of this *Plasmodium spp.* clade, which lacked the Laveranian *var*s. Indeed, the initial study found that VIR proteins were recognised by only some patient sera and localised to the surface of the few iRBCs that expressed them (del Portillo et al., 2001). These behaviours suggest that only subsets of infections express given members of the PIR family, important for antigenic switching, and furthermore that they localise to where the immune system would encounter them. In the rodent malaria *P. yoelii* it was observed that immune responses changed the *pir*s being transcribed. Pre-exposed mice altered the restriction fragment length polymorphism profile of *pir* RNA products in a secondary infection, compared to primary infection in the same mice strain and secondary infection of immunocompromised $RAG^{-/-}$ mice (Cunningham et al., 2005). Furthermore, transgenic over-expression of a VIR protein in *P. falciparum* iRBCs demonstrated binding to ICAM1 on spleen fibroblasts, so that selected PIR/VIR may be involved in cell-to-cell binding and sequestration (Bernabeu et al., 2012; Fernandez-Becerra et al., 2020).

A direct link between *pir* genes and any given function or aspect of infection, including immune evasion, is currently lacking. This is partly because genetic knock-out studies suffer from a difficulty in choosing targets from the large number of *pir* genes that exist. Large genetic ablation screens, such as PlasmoGEM, often experience limited success with variable genes like multigene families, leading to both few results and the lack of existing genetic tools pertaining to these groups (Bushell et al., 2017). There have been suggestions that the *pir*s may be involved in immune modulation. A *P. falciparum* RIFIN

has been shown to mimic human LILRB1 and downregulate NK cell activation (Saito et al., 2017), although, as mentioned above, a shared ancestry of RIFINs with PIR proteins is unlikely. Some *pir*s of *P. knowlesi* were observed to share 50% of their sequence the T-cell protein CD99, which could represent molecular mimicry in action (Pain et al., 2008).

The answer likely will not be particularly clear, as evidence is accruing that PIR proteins have different functions. For instance, their localisation within the iRBC varies between the host or parasite cytoplasm, iRBC surface, or in the parasitophorous vacuole. As identified by immunofluorescence in *P. c. chabaudi* AS*, P. berghei* and *P. yoelii*, this depends on the PIR protein (Fougère et al., 2016; Yam et al., 2016), the stage of the asexual cycle (Omelianczyk et al., 2020), and even on the method of transmission (Giorgalli et al., 2022). Giorgalli and colleagues used pan-sub-family antibodies to identify the localisation of S7 and L1 *pir*s in *P. c. chabaudi* AS, comparing mosquito-transmitted (MT) and serially blood-passaged (SBP) infections (discussed further in chapter 1.6.1) at their acute and chronic stages. The MT iRBCs showed mainly parasitophorous vacuolar membrane (PVM) expression of S7 PIRs during the acute infection, however in the recrudescing infection many more S7s were exported into the host cytoplasm and were also membrane-associated. In SBP infections the S7s were present in the iRBC cytoplasm throughout, while the L1s showed fewer changes between different parasite sources, remaining in the PVM. *In vitro* grown merozoites showed S7 and L1 expression to differing amounts (80% positive for S7s, and 95% positive for L1s) and in slightly different cytoplasmic structures (S7s in one focus and L1s in multiple cytoplasmic dots). Staining of unfixed cells (which are not permeable to antibody) was always negative, suggesting that the PIRs analysed were not exposed to the exterior of the host cell. Despite this, proteomic analysis across iRBC fractions showed that many PIRs, including L1s and an S7, could be found in the membrane fraction. A separate proteomic study of *P. berghei* that used surface shearing to distinguish intracellular from surface proteins also suggested that only some PIRs are found on the iRBC surface (Pasini et al., 2013). In summary, since the proteins are not always located at the surface of iRBCs, only some of them could be involved in cell surface interactions with the immune system.

A further major difference from variant antigen protein families such as the *var*/PfEMP1 is that transcription of *pir*s does not appear to be monoallelic, as revealed by single-cell

RNA sequencing and immunofluorescence of multiple PIRs (Cunningham et al., 2009; Reid et al., 2018). Selection of *P. yoelii* iRBC that express a given *pir* gene also show consistent expression of others, revealing that epigenetic control is certainly involved in regulating transcription but the production of one *pir*'s transcripts does not restrict the production of all others (Omelianczyk et al., 2020). An alternative hypothesis is that the PIR proteins could form an immunological smokescreen, blinding the immune system with an abundance of unimportant antigen, however this is yet to be tested. The evidence taken together presents an inconclusive picture, however it is likely that only some, if any, of the PIRs play a role in direct immune interactions.

1.6 High-throughput RNA sequencing has shed light on *pir* transcription across the life cycle

In the era of high-throughput sequencing, a new powerful avenue of understanding *pir* behaviour was opened. Before, the microarrays used to quantify gene expression frequently omitted the highly variable multigene families. Now that the total transcript content of cells could be converted to cDNA, amplified, and sequenced in its totality, these genes were no longer excluded and could be globally analysed. The muddy understanding of the role of this multigene family in *Plasmodium sp.* is reflected in transcriptomic studies of the *pir*s. Despite this, numerous investigators have shown that the data can be examined piece-by-piece to find consistent patterns of behaviour.

1.6.1 *P. c. chabaudi* AS transcriptome analysis suggests a role of *pir*s in pathogenicity

Perhaps the closest insights into function have come from *P. c. chabaudi* AS RNAseq analysis. A link between the *pir* genes and virulence of *P. c. chabaudi* AS infection was found by comparing two different modes of inoculation; the natural method of mosquito transmission (MT) and serial blood passaging (SBP) of infected samples between mice (Spence et al., 2013). When mice were infected by mosquito bite, this gave rise to a low-grade infection and eventual clearance by the host (although the parasites recrudesce later in the chronic phase of infection). Transferring the infected blood between naïve mice lead to a steady increase in virulence of infection with each iteration (Spence et al., 2013; Wéry, 1967). These more dangerous infections were characterised by high parasitaemias, potent hypothermia and anaemia, cachexia and liver damage. This

phenotype could be returned to ‘normal’, however, by transmitting these parasites through a mosquito again. This observed reversion/moderation of virulence through the vector could potentially be the result of returning to a hedge-betting strategy for new, unknown environments, after adapting to the immune environment of host mice. Notably, there were major changes in the *pir* RNAseq repertoire between MT and SBP infections. Although around 10% of the transcriptome showed significant changes between the transmission populations, the *pir*s were enriched among the differentially expressed genes. The MT parasites demonstrated a diverse and broad *pir* RNAseq profile in comparison to the SBP parasites. In the latter, one *pir* was highly upregulated to a major degree, suggesting selection of this specific *pir* for maximal fitness.

P. c. chabaudi AS was used to highlight another association between the *pir*s and pathogenicity, as a similar pattern to MT vs SBP infections was observed when acute and chronic infections were compared (Brugat et al., 2017). When acute or chronic infections of mice were passaged into naïve mice, the transferred acute parasites showed a broad, diverse *pir* repertoire, but *pir* expression of the transferred chronic parasites was dominated by a much smaller subset. Again, this change was in tandem with an increased virulence of chronic-derived parasites in naïve mice (the chronic infection in the original mouse, which by now had developed immunity, causes minor symptoms). Another notable observation was how these changes corresponded to sub-families and chromosomal locations of the genes involved. It was found that the *pir*s upregulated in acute/chronic infections tended to group in given loci. These genetic clusters of co-expressed *pir*s were termed acute-associated *pir* loci (AAPLs) and chronic-associated *pir* loci (ChAPLs) depending on which stage of infection they were primarily associated with. The loci were enriched for given sub-families, with S7s found to dominate the AAPLs and L1s the ChAPLs, highlighting the potential use of the groupings in identifying functional distinctions. A similar signal, albeit a much-reduced change, was also observed in *P. berghei* samples from a chronic model of Norwegian brown rats. Intriguingly quite the opposite effect is found in chronic *P. coatneyi* infections of rhesus macaques, with the chronic *pir* profile showing upregulation overall (but variable between genes) (Cordy et al., 2019). There is a consistent pattern of the *pir*s changing alongside an adaptation of the parasites to a given immune environment, which in the case of *P. c. chabaudi* AS exacerbated the pathogenicity of infection in non-immune mice.

1.6.2 RNAseq data from other species demonstrates patterns among *pir* transcription

The rodent-infecting *P. berghei* remains the dominant model in mice, and it has had its transcriptome sequenced many times in different contexts. The earliest such study, Otto et al, 2014, covered the asexual/sexual blood stages as well as mosquito stages, and they observed that while many *pir*s were transcribed among the former samples, the latter generally showed little-to-no *pir*s. This is a pattern that is frequently noted when the transcriptome of *P. berghei* mammalian stages are compared to those that develop in the mosquito, suggesting that the *pir*s do not play an important role in the mosquito. This is one such example of how transcriptomic studies can enlighten broad functional understanding. It was not long before methods had enhanced resolution and one could perform RNAseq of individual cells (single cell RNAseq; scRNAseq). scRNAseq was applied to *P. berghei* sexual cells, showing for the first time the difference in *pir* transcription between male and female gametocytes. The female cells seldom transcribed any *pir*s, while the male gametocytes showed highly variable but much-expanded transcription (Reid et al., 2018).

P. vivax RNAseq was used to quantify the *pir* transcriptome across the asexual blood cycle, using synchronised cells cultured *ex vivo* (Zhu et al., 2016). As with *P. berghei* a broad swathe of *pir*s were transcribed across the entire cycle, however there were extremes of transcription observed at the end of schizogony/beginning of the ring stages. Hence even within the life cycle stages that transcribe high numbers of *pir*s, there are differences to be found with the timing of development. These could suggest the importance of *pir*s or sub-divisions of *pir*s primarily at certain points of the asexual cycle. By comparing two different patient isolates of the parasites, they also observed a group of *pir*s transcribed highly in both isolates, as well as groups of *pir*s that were only transcribed in one isolate and not the other, hinting at the diversity of *pir* repertoires between infection populations and/or hosts.

Despite the insights these studies have garnered about when *pir*s may be necessary in *Plasmodium sp.* infection, seldom are the *pir*s incorporated into the analysis of the numerous other studies that extracted the transcriptomes of these species. These studies present rich data which have not yet been mined for their *pir* potential.

1.7 The aims and hypothesis of this thesis

With this multitude of studies, an opportunity arises of not only analysing the change of *pir* transcription across many stages of the life cycle, but also assessing the consistency of results. For a multigene family, results can be starkly different between experiments, however where consistency emerges, we can identify reproducible moments when *pir* mRNA is likely to be required by malaria parasites. This thesis uses published (*P. berghei* and *P. vivax*) and novel (*P. c. chabaudi* AS) data to investigate an open question about *pir* genes in infection across species and life cycle stages; Which *pir* genes, and *pir* sub-families, are consistently associated with stages of *Plasmodium sp.* development? By understanding when *pir*s are transcribed across species it will be possible for future investigators to hone their studies to the time of the life cycle most relevant to PIR function. These observations may also improve hypotheses on the functions of PIRs/PIR clades, as the timing of transcription will be linked to given processes that the parasite undertakes to survive, such as proliferative stages or stages exposed to adaptive immunity.

With multiple species included, I can also compare between species, and I hypothesise that *pir* gene transcriptional patterns will be conserved between them. Furthermore, I suggest that, although *pir* sub-families do not have direct phylogenetic relations between the rodent and the primate malaria clades, there will be equivalent sub-families identifiable through their expression profiles. Finding conserved transcriptional patterns of *pir*s between species will suggest conserved functions of the genes between rodent models and human infections.

Another important resolve of this thesis is to identify targets for future experimentation, such as localisation studies and gene KOs. Multigene families make initiating these investigations challenging as there can be thousands of members to sift through for the choice to target. Many may be redundant, leading to experiments with few insightful results. The genes which show a clear and consistent signal for overall development or development at certain stages make clearer, delineated options for studying a phenotype. Even sub-family associations present a subset of targets for a mid-throughput screen.

In summary, my objectives are to:

1. Use ten *P. berghei* published RNAseq datasets from across the life cycle to find consistent associations of *pir* genes and sub-families with given stages of development. (Chapter 3)
2. Identify constitutively transcribed *pir*s as prime targets for future study (Chapter 3)
3. Use a novel dataset of *P. c. chabaudi* AS RNAseq taken across the asexual life cycle, in different methods of transmission, to gain greater resolution of this stage of development and how *pir*s associate with each time point. (Chapter 4)
4. Compare and contrast the transcription of *pir* genes between different rodent-infective species, and between rodent and human-infective species. (Chapters 4 and 5)
5. Use published *P. vivax* RNAseq datasets for *de novo* transcriptomic assembly to ascertain the diversity and amount of *pir* transcription between samples, conditions, and life cycle stages. (Chapter 5)

Chapter 2. Materials and Methods

2.1 *P. c. chabaudi* AS culturing and RNAseq extraction

2.1.1 Attribution of labour

Note that this section of the Materials and Methods was entirely performed by others, as described in the text.

2.1.2 Mice

Infections of *P. c. chabaudi* AS were maintained in female C57Bl/6 mice of 6-8 weeks of age. These mice were reared in a specific pathogen free environment at The Francis Crick Institute Mill Hill site (National Institute of Medical Research), at temperatures of 20–22 °C, and with access to mouse breeder diet/water at all times. They were raised in normal light (light 07.00–19.00, dark 19.00–07.00 GMT) and reverse light (light 19.00–07.00, dark 07.00–19.00 GMT) conditions, the latter environment for minimum a week before infections were performed. Care of mice was performed by Audrey Vandomme, Thibaut Brugat, other members of the Langhorne Lab and staff of The Francis Crick Institute Biological Research Facility. The study conformed to the UK Animals (Scientific Procedures) Act 1986 (operating under the Home Office licences 80/2538 and 70/8326) and received ethical confirmation from The Francis Crick Institute Ethical Committee.

2.1.3 Mosquitoes

Anopheles stephensi mosquitoes were bred at The Francis Crick Institute Mill Hill site. They were kept at 26°C and fed with 8% (w/v) D-Fructose (Sigma, cat. no. F0127) and 0.05% (w/v) para-aminobenzoic acid (Sigma, cat. no. A9878) solution in dH_2O (sterilised through filtration) which was always available and replaced every two days. To instigate egg production the insects were provided with defibrinated horse blood (E & O Laboratories Ltd) from a Hemotek membrane feeder. Egg bowls were provided as 0.3g/L sea salt solution with moist filter paper partially submerged. Eggs were collected and placed in larger trays, and as the eggs hatched and larvae grew the colony was split into further trays. Larvae were fed Liquifry No. 1 (Interpet) and pool pellets (Dr Clarke's pond foods). Mosquitoes were reared by Prisca Hill, Irene Tumwine, Audrey Vandomme, Thibaut Brugat, and other members of the Langhorne lab.

2.1.4 Parasite maintenance

All *P. c. chabaudi* AS strain used in these experiments was originally sourced from a clonal isolate provided by Professor David Walliker, University of Edinburgh, UK. The stock was maintained and boosted by infecting mice by i.p. injection of iRBCs and passaging inoculum between naïve individuals for new cryopreserved stabilates to be frozen in liquid nitrogen (SBP). For infection studies or maintaining parasite stock SBP is only iterated four times between mice (the cryopreserved stock is itself a product of SBP so overall the parasite has been through up to six passages since last mosquito transmission). Mosquito transmitted (MT) infections of *P. c. chabaudi* AS used *Anopheles stephensi* insects and was performed as described in (Spence et al., 2012). Briefly, SBP-infected mice (at least two weeks post infection) with gametocytaemia above a threshold of 0.1% (i.e. with $\geq$ 0.1% i.e. of all blood cells infected with gametocytes), as determined by light microscopy of Giemsa-stained thin blood films (Giemsa - BDH (VWR), cat. no. 352603R), were terminally anaesthetised and placed onto the mosquito cages with access for the insects to bite. After four days the mosquitoes were provided with an egg bowl, and they were given a second feed at six days to enrich *P. c. chabaudi* AS infection. Recently mosquito transmitted (RMT) infections were obtained by infecting naïve mice with iRBCs from MT infections. These singly-passaged infections have been shown to retain the transcriptional and phenotypic characteristics of MT parasites, but the mice can be given a consistent iRBC administration instead of variable sporozoite numbers from mosquito bite. All iRBC inoculum were injected i.p. at a constant number of 10^5 infected cells. The raising of parasite in mice, blood collection and parasitaemia counts were performed by Audrey Vandomme, Thibaut Brugat, and other members of the Langhorne Lab.

2.1.5 RNA extraction and sequencing

Blood for RNA was removed from 3 exsanguinated mice at each timepoint of infection (02:00, 05:00, 08:00, 11:00, 14:00, 17:00, 20:00, 23:00) seven days from inoculation. The blood was filtered through Plasmodipur (Euro Proxima) to deplete leukocytes. Host RNA was depleted by saponin lysis and centrifugation. RNA extraction itself was performed by the protocol of (Spence et al., 2013) in which the purified parasites were mixed with Trizol (Invitrogen 15596026), frozen at -80°C, and later resuspended in water for library preparation. The samples were converted into 150-350 bp fragment Illumina

TruSeq libraries through 12 cycles of PCR amplification, then for sequencing they were tagged, pooled and ran on two Illumina HiSeq2500 lanes. Each sample was present on both lanes and sequenced as 100bp reads. RNA extraction and library preparation was performed by Audrey Vandomme, Thibaut Brugat, other members of the Langhorne Lab, and staff of The Francis Crick Institute's High-Throughput Sequencing Science Technology Platform.

2.2 RNAseq analysis of the *P. c. chabaudi* AS and *P. berghei* data

2.2.1 Dataset download and processing

The *P. c. chabaudi* AS data and the *P. berghei* data from Witmer et al., 2020, and Ukegbu and Giorgalli et al., 2020, were transferred manually from lab storage. All other *P. berghei* RNAseq data was downloaded from the Sequence Read Archive (SRA) with SRA-Tools (SRA_Toolkit_Development_Team, n.d.). The cut-off date for inclusion in the study was December 31st, 2020 for *P. berghei*. The technical replicates among these data were concatenated before processing. For analysis the rna-seq pipeline v1.4.2 (from nf-core, with Nextflow workflow software v20.07.1) was used (di Tommaso et al., 2017; Ewels et al., 2020). Paired-end and single-end samples were analysed with identical but separate workflows. As part of this workflow FastQC v0.11.8 (Andrews, n.d.) and MultiQC v1.7 (Ewels et al., 2016) were used for quality control, Trim Galore! v.0.6.4 (Krueger, 2015) removed adapter sequences from the RNAseq data, HISAT v.2.1.0 (Kim et al., 2015) aligned the sequences to the reference genomes, and featurecounts v1.6.4 (Liao et al., 2014) functioned to calculate the gene counts from the alignment. The configuration profile used for Nextflow was a locally maintained profile for The Francis Crick Institute (https://github.com/nf-core/configs/blob/master/docs/crick.md). The gene annotations, including for the different *pir*s, and reference genomes (Brugat et al., 2017; Fougère et al., 2016) for featurecounts gene counting were taken from PlasmoDB v48 (a resource supported by GeneDB) (Aurrecoechea et al., 2009; Logan-Klumpler et al., 2012). For ChAPL and AAPL designation I followed Brugat et al., 2017. I excluded the annotated *pir*s *PBANKA_0524600* and *PCHAS_0524800* due to their two-exon structure which differs from most other pirs. The gene counts were further analysed in R v4.0.2 (R Core Team, 2018). They were converted into TPM by normalizing for gene length and library size, a

metric chosen over RPKM/FPKM (RNA/Fragments Per Kilobase of exon per Million reads mapped) as the latter varies in its total between samples (Wagner et al., 2012).

2.2.2 Data filtering and analysis in R

The similarity and correlation between the samples' global TPM values were investigated using Principal Components Analysis (PCA – with the inbuilt prcomp function in R with 'center' and 'scale' set to TRUE), hierarchical clustering (R's hclust function) and Pearson correlation. It was decided based on this that some of the times and conditions of the original RNAseq samples should be combined due to transcriptional similarity, and others were removed due to poor quality or dissimilarity to other samples. The samples combined were from nearby timepoints and minor condition changes; liver stage samples at 4h post-infection or earlier (Liver.pre4h), liver stage samples at 48h post infection or later (Liver.post48h), liver stages from the same timepoints but different cell lines, the chronic and acute samples from Brugat et al., 2017 (Asexual.Mixed), the different times of ookinete development from Otto et al., 2014 (Ookinete), the 'activation' state of gametocytes from Ukegbu and Giorgalli et al., 2020 (Gametocytes), and 3h post-uptake and earlier blood meal samples from Ukegbu and Giorgalli et al., 2020. (BloodMeal-Pre3h). The 6h Liver samples from Caldelari et al., 2019 (E06), were removed due to low read counts, as performed in the original publication. A 2h Liver sample (SRR11142819) from Toro-Moreno et al., 2020 (E08), and two schizont samples (SRR3437888 and SRR3437912) from Modryzynska et al., 2017 (E03), were removed due to being different to other samples from the same experiment by hierarchical clustering of the entire transcriptome (see Figure 3.2).

2.2.3 Statistical analysis of differential expression in R

I treated each biological replicate from each experiment as an independent sample. For comparison of gene set transcription (such as the *pir*s, *pir* sub-families and *pir* loci) I used Qusage v2.22.0 (Meng et al., 2019) with the TPM data and $2x10^{18}$ iterations. With *P. c. chabaudi* AS I only included genes in a gene set if they were transcribed above 10 TPM in at least one timepoint sample, using the re-normalized data without pseudogene *PCHAS_0420800*. The sub-families L2 and the three unannotated *P. c. chabaudi* AS *pir*s were not included as they had only three members or fewer, which Qusage recommends not including. I made pairwise contrasts between consecutive timepoints

within recently mosquito transmitted (RMT) and serially blood passaged (SBP) (including log fold changes between 02h and 23h, as this is when the cycle theoretically restarts), and between the same timepoints of RMT and SBP. The L-form statistical comparisons between the schizonts and other asexual stages, for *P. c. chabaudi* AS and *P. berghei*, were conducted using the prop.test function in R, as Qusage cannot handle different genomes.

For statistical comparison of the transcription of individual genes between samples I used DESeq2 v1.28.1 (Love et al., 2014). Log fold change shrinkage was applied with the ashr model (M. Stephens, 2016) so the genes with low transcription have less of an effect on the results. Differential expression was considered statistically significant if the adjusted p-value (multiple testing correction by the false discovery rate (Benjamini & Hochberg, 1995)) was less than 0.05.

2.2.4 Deconvolution of *P. c. chabaudi* AS bulk RNAseq

This was performed by Adam Reid following the protocol of Aunin et al., 2020. Briefly, the single-cell transcriptomes from the Malaria Cell Atlas were used to produce pseudobulk samples of circulatory stages of the parasite which were supplied as the reference for CIBERSORT v1.06 (default settings) (Newman et al., 2015).

2.3 Phylogenetic analysis of the *pir* 'ancestral' sequences

To find orthologs of the ancestral *pir* gene (Frech & Chen, 2013) 'transformation by orthology' search (which invokes OrthoMCL (L. Li et al., 2003)) was performed on PlasmoDB v48 using *PBANKA_0100500*. The ancestral gene for *P. c. chabaudi* CB was extracted using a BLAST search of the genome (it was 100% identical to the *P. c. chabaudi* AS gene) (Lin et al., 2018). This found nineteen orthologs across the Plasmodia. These were uploaded to the Seaview platform (Gouy et al., 2010) and aligned using MUSCLE v3.8.31 (Edgar, 2004) and a phylogenetic tree was built with PhyML v3.1 (Guindon et al., 2010) with LG model 100 bootstrap replicates (otherwise default settings). The tree was trimmed by removing sequences from strains of the same species for simplification. To design and visualise the tree, in R, the packages treeio v.1.12.0 (Wang et al., 2020), ggtree v2.2.1 (Yu et al., 2017), and ape v5.4 (Paradis & Schliep, 2019) were used.

2.3.1 Transcriptional comparisons of the ancestral *pir* across the Plasmodia

Using these orthologs, transcription data from published studies (see Table 3.2) was analysed to see in which percentage of the global transcriptome the ancestral *pir* was found in e.g. was it among the 10% highest transcribed genes? The published TPM/RPKM files from these studies were used to calculate this.

.

2.4 *De novo* transcriptomic analysis of *P. vivax* RNAseq data Methods

2.4.1 *P. vivax* data download and adapter trimming

Datasets were downloaded using the FetchNGS pipeline v1.5 available from nf-core using Nextflow v21.10.3 and Singularity v3.6.4. Default settings were used except for '--force_sratools_download' to use SRA-toolkit for the download, and an additional custom config file (https://github.com/timslittle/Thesis_Github_221023/blob/main/custom.config) supplied via '-c' to permit the SRA-toolkit function 'prefetch' to use larger files than default. As before, the configuration profile used for Nextflow was a locally maintained Crick profile (https://github.com/nf-core/configs/blob/master/docs/crick.md). Samples were quality filtered and adapter trimmed using fastp v0.20.0 (Chen et al., 2018) with the settings "--compression 9 --detect_adapter_for_pe" for maximum file compression and automatic adapter sequence detection. At this point files were concatenated together if they constituted the same isolate of *P. vivax*, to maximise the amount of information for the transcriptome assemblers, but minimise the erroneous assembly of transcripts from stitching reads of different sources. One dataset which was considered for this study, Roth et al., 2018, was excluded because the source of the parasite isolates was not clear from the methods. See Chapter 7 - Appendix for the full list of files concatenated together and the rationale for this.

2.4.2 *P. c. chabaudi* AS RNAseq pre-processing, *de novo* transcriptome assembly and analysis

The *P. c. chabaudi* AS RMT 14h RNAseq samples were aligned to the GRCm39 *Mus musculus* genome (Genome Reference Consortium: https://www.ncbi.nlm.nih.gov/grc) to remove reads which aligned. HISAT2 v 2.1.0 was the aligner used with parameters '-q -x --un-conc-gz', with the last parameter saving the sequences which do not align concordantly to the reference (Kim et al., 2015). Bbduk (BBMap v36.20) was then employed to perform k-mer depletion for rRNA reads (Bushnell et al., 2017), using the SILVA rRNA database (138.1) (Quast et al., 2013), and the following parameters 'k=25 mcf=0.9'.

It is upon this processed 14h RMT sample and the concatenated full *P. c. chabaudi* AS RMT and SBP dataset that *de novo* transcriptome assembly was performed. Trinity v2.2.0 was used with the parameters '--seqType fq --max_memory 128G' (Haas et al., 2013). SPAdes v3.9.0 was used with the parameter '-m 128', with the python script rnaspades.py used for rna-spades assemblies, and spades.py and the '--sc' parameter for sc-spades assemblies (Bankevich et al., 2012; Bushmanova et al., 2019). CD-hit-est v4.8.1 redundancy removal was performed on the 14h RMT samples, with parameters '-n 10' and '-c' set to 0.9, 0.95 or 0.99 (Fu et al., 2012; W. Li & Godzik, 2006).

To evaluate how well the output captured the *pir* transcripts of *P. c. chabaudi* AS I used BLAST+ v2.9.0 to align the *de novo* transcriptomes to PlasmoDB v53 annotated *P. c. chabaudi* AS *pir* sequences (Altschul et al., 1997; Aurrecoechea et al., 2009). The parameters '-evalue 1e-20 -num_threads 6 -max_target_seqs 50 -outfmt "6 std qcovs qcovhsp slen nident"' were used to filter the results and increase the number of reported output statistics (mainly for the 'qcovhsp' metric). The results from BLAST were finally analysed in R (R Core Team, 2018).

2.4.3 *P. vivax* RNAseq pre-processing and *de novo* transcriptome assembly

The *P. vivax* RNAseq datasets (both the concatenated RNAseq files and individual files) were aligned similar to above but using a concatenation of the *Mus musculus* genome (due to the mouse fibroblasts in the Gural et al, 2018 system), the *Homo sapiens* genome GRCh38 (Schneider et al., 2017), *Anopheles dirus* WRAIR2 mosquito AdirW1

genome (due to the mosquito-sourced samples of Boonkaew et al., 2020) (Amos et al., 2022; Neafsey et al., 2015), as well as the primate *Saimiri boliviensis* SaiBol1 genome (Chiou et al., 2011) and *Aotus nancymaae* GCA_000952055.2 genome (Babb et al., 2011) (the monkeys used in Gunalan et al., 2019). The HISAT2 alignment removal and bbduk k-mer depletion were performed as described above except with the relevant parameters altered for the single-end and paired-end stranded *P. vivax* samples: for the stranded libraries of Kim et al., 2017 and 2019 samples HISAT2 was performed with '--rna-strandness FR'; for the single-end libraries of Muller et al., 2019, and Gural et al., 2018, HISAT2 was performed with '--un-gz' instead of '--un-conc-gz'.

The fastp-processed, HISAT2-filtered and bbduk-depleted RNAseq datasets were then used for building *de novo* assemblies with sc-spades (SPAdes v3.9.0, parameters '-m 128 --sc'). Only the concatenated isolate files were used for this purpose unless an isolate was made up of only one RNAseq dataset and hence it was not concatenated. CD-hit-est v4.8.1 was used with a redundancy threshold of 95%: '-c 0.95 -n 10'.

2.4.4 *P. vivax pir* transcript detection

The sequence assemblies were processed using getorf (EMBOSS v6.6.0, parameters '-minsize 300 -table 0 -find 1') to detect and translate open reading frames, with a minimum length of 300bp (100aa), between a standard eukaryotic start and a stop codon (or alternative eukaryotic initiation codons) (Rice et al., 2000). These ORF files were used to calculate BUSCO scores with BUSCO v4.0.5 and settings '-m proteins' and 'apicomplexa_odb10' and 'plasmodium_odb10' as references (Manni et al., 2021; Simão et al., 2015). HMMer v3.3.2 is then used to identify the sequences which resemble the Plasmodium_Vir Pfam model (PF05795) (Mistry et al., 2021), then matches with an E-value of less than 1e-3 were extracted (Eddy, 2011). The ORFs are then filtered to get those at least 200aa long using seqkit 'seq -m 200' (Shen et al., 2016).

Transcriptional quantification of the samples was then used to both detect which of the proposed *pir* transcripts have RNAseq reads map back to them, and to evaluate how the *pir*s are being transcribed. All the original biological replicates (after fastp, HISAT2 and bbduk processing) were aligned to their corresponding assemblies using the transcript-aware aligner Salmon v1.9.0 (Patro et al., 2017). The parameters '-p 8 --

validateMappings -l A' were used for all alignments, with '-l SR' specified instead of '-l A' for the stranded libraries. Salmon calculates TPM automatically, so these results were used directly to filter the HMMer ORF outputs for the PIR sequences with evidence of transcription (greater than or equal to 1 TPM in at least one sample).

2.4.5 Creating networks of PIRs and sub-family assignment

To create networks of the PIR ORFs and work out which sub-family they belong to/how well existing definitions describe these PIRs, I downloaded the already published PIR sequences from *P. vivax* strains P01 and Sal1, as well as the *de novo* PIRs from Zhu et al., 2016 (denoted as Zhu16). For P01 and Sal1 a PlasmoDB v56 search was performed for "PIR protein" in 'Product Description' as well as a Pfam search for Plasmodium_Vir. For Zhu16 the original *de novo* gene sequences were processed through getorf in the same manner to the original *de novo* sequences of this thesis, which gave 116 ORFs from 102 mRNAs, out of the original 179 total mRNAs. These 'known' PIRs were combined with the *de novo* PIRs and BLASTp v2.9.0 was performed between this dataset and itself ('-evalue 1e-3 -num_threads 6 -max_target_seqs 500 -outfmt "6 std qcovs qcovhsp slen nident"').

The results were minimally filtered for at least 10% identity alignments and 20aa long alignments (and removing all same-to-same alignments) before performing MCL clustering with varying inflation values (influence the number of clusters) of 1.2, 1.4, 1.8, 2, 2.5, 3, 4 and 6 (mcxload parameters: '--stream-mirror --stream-neg-log10 -stream-tf 'ceil(200)''; mcl and mcxdump parameters all default) (van Dongen & Abreu-Goodger, 2012). The networks were further filtered for only alignments with at least 20, 30, 40, 50, 60, 70, 80, 90% identity and 50, 75, 100, 150, 200, 250 and 300aa length in order to assess the best networks. As reported in the results, the best compromise between minimal exclusion of PIRs from the network, minimal total number of clusters and minimal mixing of Sal1 sub-family sequences was found with 30% identity and 25aa long BLAST alignments, and an MCL inflation value of 1.4. Based on the MCL clusters sub-families were assigned based on the majority Sal1 sub-family sequence in the cluster. If a cluster had no assigned Sal1 sub-family sequence (note that there may be unassigned Sal1 sequence present) then it was Unassigned. These were then numbered if the assigned name was not unique.

The network was visualised using Gephi (Bastian et al., 2009) and the OpenOrd layout algorithm (Martin et al., 2011) with 10% Liquid stage, 50% Expansion stage, 5% Cooldown stage and no Crunch or Simmer stage of the simulated annealing process, 0.9 edge cut, 7 threads, 1000 iterations, 0.2 fixed time and a random seed of 234, followed by briefly running the Noverlap algorithm to reduce overlapping nodes with speed set to 3, ratio 0.5 and margin 5.

2.5 Other packages used in R throughout analysis

R packages used in manipulating and analysing this data include dplyr (Wickham et al., 2019), data.table (Dowle & Srinivasan, 2019), reshape2 (Wickham, 2007), stringr (Wickham, 2019), tibble (Müller & Wickham, 2019), readr (Wickham et al., 2018), and readxl (Wickham & Bryan, 2019).

Visualization was achieved using ggplot2 (Wickham, 2016), RColorBrewer (Neuwirth, 2014), corrplot (Wei & Simko, 2017), plotly (Sievert, 2020), viridis (Garnier, 2018), circlize (Gu et al., 2014), and ComplexHeatmaps (Gu et al., 2016).

2.6 Availability of R code

Code used in R for analysis can be found at https://github.com/timslittle/Thesis_Github_221023.

Chapter 3. Results - Analysis of *P. berghei pir* transcription across the life cycle

3.1 Introduction

Major indicators for the role of a *Plasmodium* gene can come from its association with stages of the life cycle. For example, the gene *p47* is upregulated in (female) gametocytes and is involved in protection of the *Plasmodium* ookinete from the innate immunity of the mosquito (Molina-Cruz et al., 2013; Ukegbu et al., 2017). The *Plasmodium falciparum var* genes are (almost) exclusively transcribed in the asexual blood stages, which is because of the function of the encoded PfEMP1 proteins in facilitating the binding of infected red blood cells (iRBCs) to endothelial tissue (Su et al., 1995). Transcriptional analysis of the *pir* gene family, the subject of this thesis, has also been important in improving our understanding of these genes. By utilising RNAseq data from the rodent malaria parasite *P. berghei* we know that these genes are likely to be important in the asexual blood stages compared to potential unimportance in the mosquito stages (Otto et al., 2014). Additionally, *P. c. chabaudi* AS transcriptomes have revealed a link between the genes and both virulence and chronicity of infection (as elaborated upon in Chapter 1) (Brugat et al., 2017; Spence et al., 2013). Single-cell RNAseq has been used to expand the coverage of the *P. berghei* transcriptome to the entire life cycle, reinforcing previous conclusions, revealing the dearth of liver-stage transcription, and the disparity between the different sexes of gametocytes (Howick et al., 2019; Reid et al., 2018). However, the resolution of single-cell technology comes with the cost of a loss of gene sensitivity. Current single-cell technology can only reliably capture genes transcribed at higher levels, and *pir* genes are often under-detected (Howick et al., 2019; Kharchenko et al., 2014). Hence it is important also to have a full life-cycle analysis of this gene family using bulk RNAseq data.

Although a full life-cycle analysis of bulk RNAseq for the *pir* genes has not been performed, the data needed already exist. There are extensive RNAseq datasets covering most stages of the *P. berghei* life cycle from multiple different published studies. This parasite lends itself well to experimentation since it infects laboratory mice, and all stages of its life cycle can be easily isolated (Sinden et al., 2002; Yoeli & Hargreaves, 1974). Additionally, there are relatively few *pir* genes (134 canonical *pir*

genes in total) in the *P. berghei* genome. This compactness is an advantage for experimental investigation as it gives fewer targets to study, and there may be less redundancy in gene function between the *pir*s. The *Plasmodium berghei pir* genes have also been assigned into different sub-families based on phylogeny and sequence length (Otto et al., 2014). This includes broader groupings such as the 106 Short *pir*s (S-forms), 27 Long *pir*s (L-forms), and a single distinct clade for a gene thought to be the ancestor of the family (Frech & Chen, 2013). The L and S-forms can be further sub-divided into smaller groups. I can hence perform analysis over defined subsets of the family instead of just all the genes or individual members.

I identified ten published datasets which cover virtually the whole life cycle of *P. berghei*, including the asexual cycle, liver stages, sexual stages and the beginning/end of mosquito passage (see Figure 3.1 and Table 3.1) The *P. berghei* asexual cycle is asynchronous, meaning that extractions from blood will yield multiple stages. Therefore, most of the RNAseq data from the five different studies covering asexual stages uses mixed samples. One paper (Otto et al., 2014) used brief *in vitro* culture to synchronise the parasite growth, and to enable extraction of individual life cycle stages in the blood. Most analyses of the sexual stages similarly did not discriminate based on sex, while two studies separated male and female gametocytes for RNAseq. One experiment sampled the mosquito blood meals, which includes a mixture of asexual, sexual and ookinete stages. All the liver-stage experiments covered multiple time points along the progression from sporozoite to mature exoerythrocytic schizont.

The major problem with the comparison of RNAseq data from diverse publications is variation between the different laboratories performing the investigations. Laboratory-specific effects can be influenced by the people who performed the experiments, the exact strain and isolate of the parasite used, the different types/strains of mice, cell cultures and mosquitoes used, the batches of reagents utilised, the timings during procedures, the protocols employed, etc. These are ultimately unknowable, and can only

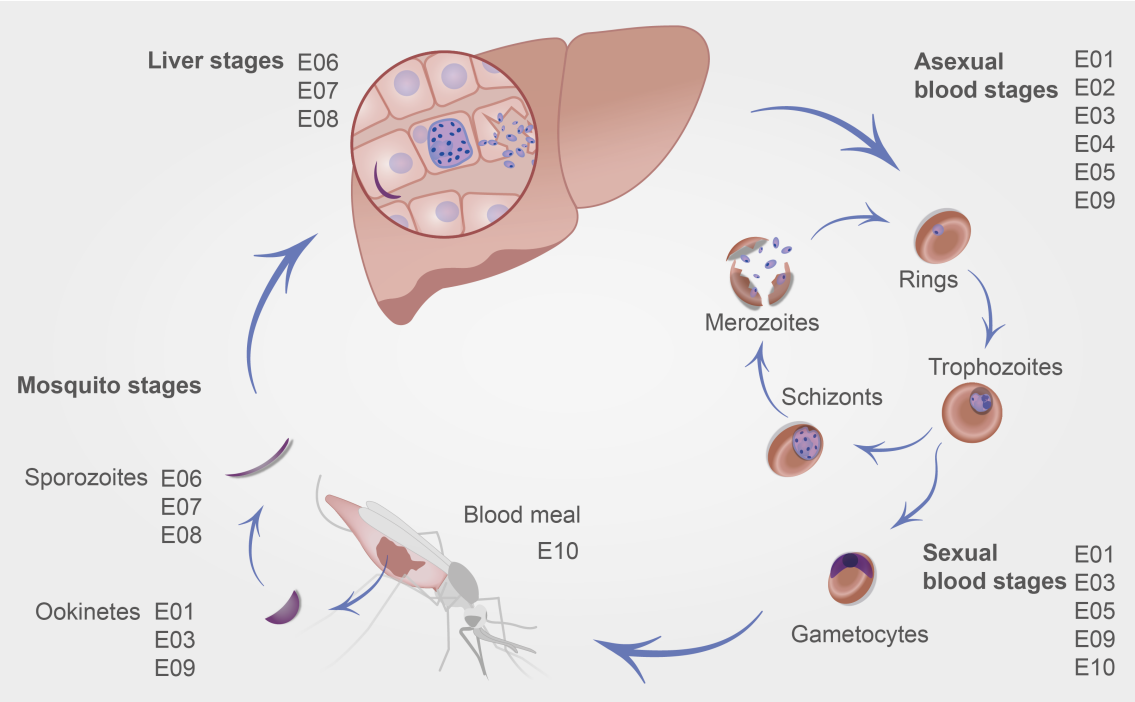


Figure 3.1 - Schematic of the life cycle stages of *P. berghei* covered by the different RNAseq publications used in this investigation.
Experiment codes as specified by Table 3.1. Designed by Ana Antonello.

be estimated by internal controls, or by comparing samples of the same stage (Risso et al., 2014). Unfortunately, this is impossible to do for all these data, as the liver-stage experiments (E06-08) have no overlapping samples with other experiments (E01-E05, E09-10). Hence, it was necessary for me to investigate the robustness of the data before any comparisons could be made.

After re-analysing the raw sequences through a common pipeline, I was able to confirm that batch effects were minor. The transcriptional analysis demonstrated consistent patterns of *pir* mRNA activity across different life cycle stages of the parasite, as well as associations of given individual genes and sub-families.

3.2 Aims

- Verify the robustness of the experimental *P. berghei* RNAseq datasets and the presence of batch effects between publications.

- Observe the amount and consistency of *P. berghei pir* transcription across the life cycle, between different stages and datasets.
- Identify individual and sub-groups of *pir*s which are enriched in given stages of the life cycle or transcribed throughout.

3.3 Results

3.3.1 RNAseq samples from different stages of the *P. berghei* life cycle robustly cluster by the parasite developmental stage

As this investigation uses RNAseq datasets from multiple different publications (conducted in different labs, by different people (see Table 3.1) there are many variables between the results which cannot be controlled for. Hence, the first step of this comparison is to test whether the life cycle stages of *P. berghei* are more similar to samples extracted from the same stage of the life cycle than samples extracted as part of the same experiment/publication. The latter case would indicate extreme batch effects and could invalidate the results of comparisons without major correction. In order to reduce the possibility of batch effects influencing the investigation as much as possible, the raw sequencing data from the original studies were re-analysed using a common bioinformatics pipeline. The resultant count data for individual genes were then further normalized as TPM (Transcript-per-million).

Initially, the data in each publication were assessed separately to look for intra-experiment outliers. To do this, hierarchical clustering was performed on the global transcriptional profile of each sample from each publication. While the vast majority of samples associate with others taken from the same stage of the life cycle, a few are noticeably different. While this could be due to capturing a distinct biological phenomenon, as there are generally few highly divergent samples, I make the assumption that this is due to error, and they should be removed from the study.

Experiment Code	Reference	Stages
E01	A comprehensive evaluation of rodent malaria parasite genomes and gene expression – Otto et al., 2014	Asexual: Rings; Trophs; Schizont. Gametocytes; Ookinetes.
E02	Antibody-independent mechanisms regulate the establishment of chronic Plasmodium infection – Brugat & Reid et al., 2017	Asexual: Mixed.
E03	A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the Plasmodium Life Cycle – Modrzynska et al., 2017	Asexual: Schizont; Gametocytes; Ookinetes.
E04	Nutrient sensing modulates malaria parasite virulence – Mancio-Silva et al., 2017	Asexual: Mixed
E05	Comparative transcriptomics of female and male gametocytes in Plasmodium berghei and the evolution of sex in alveolates – Yeoh et al., 2017	Asexual: Mixed; Gametocyte: Female; Gametocyte: Male
E06	Transcriptome analysis of Plasmodium berghei during exo-erythrocytic development – Caldelari et al., 2019	Sporozoites; Liver: 24h; Liver: post 48h (48h, 54h and 60h); Liver: Merozoites
E07	Dual RNA-seq identifies human mucosal immunity protein Mucin-13 as a hallmark of Plasmodium exoerythrocytic infection – LaMonte et al., 2019	Sporozoites; Liver: 24h; Liver: post48h (48h)
E08	RNA-Seq Analysis Illuminates the Early Stages of Plasmodium Liver Infection – Toro-Moreno et al., 2020	Sporozoites; Liver: Pre4h (2h and 4h); 12h; 18h; 24h; 36h; 48h
E09	An epigenetic map of malaria parasite development from host to vector – Witmer et al., 2020	Asexual: Mixed; Gametocyte: Female; Gametocyte: Male; Ookinetes
E10	PIMMS43 is required for malaria parasite immune evasion and sporogonic development in the mosquito vector – Ukegbu & Giorgalli et al., 2020	Blood Meal: Pre3h (1h and 3h); Blood Meal: 24h; Ookinetes

Table 3.1 - Published studies from which P. berghei RNAseq datasets were used in this chapter, including the stages of P. berghei extracted and the experimental code used to identify sample origins.

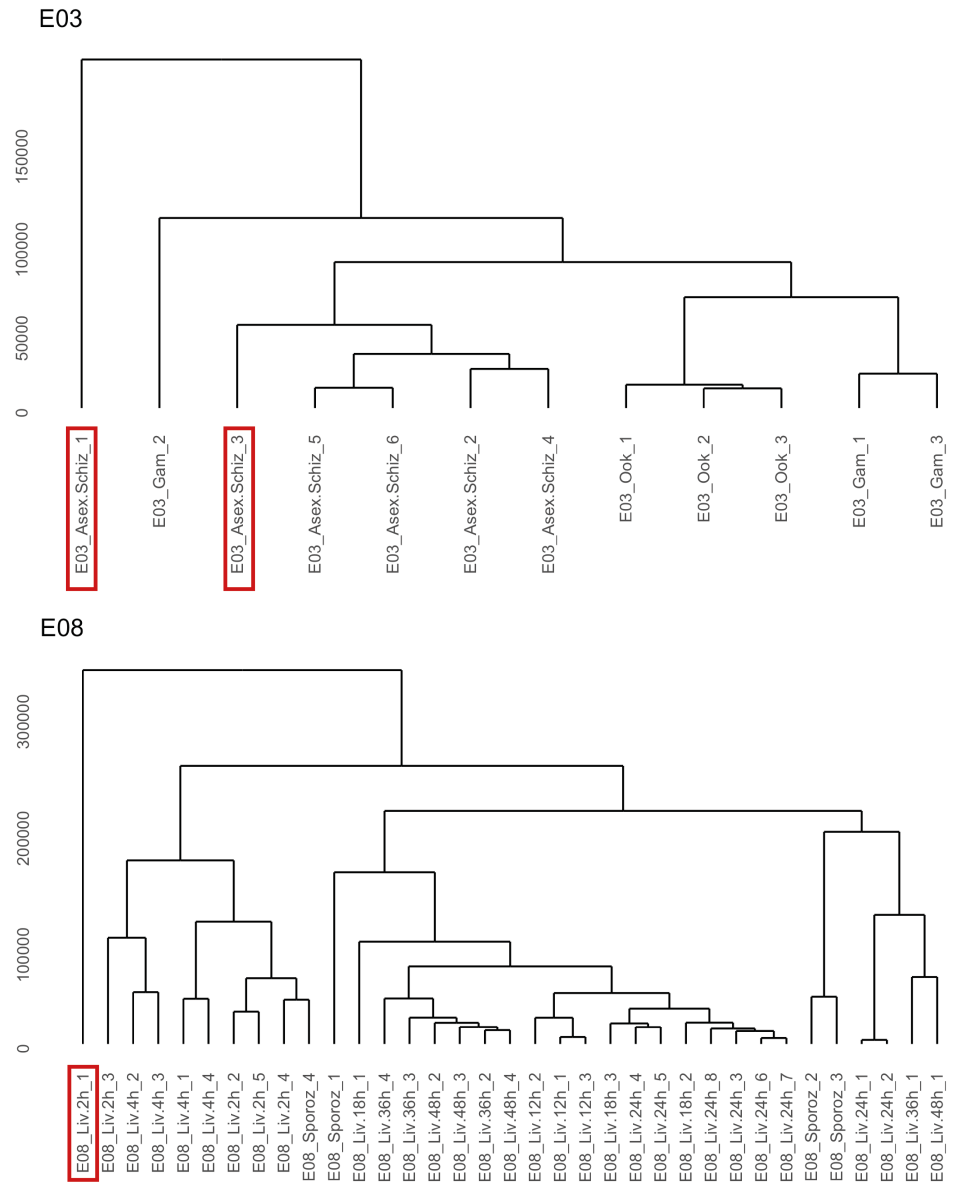


Figure 3.2 - Hierarchical clustering of the P. berghei RNAseq samples in TPM from E03 and E08.

Uses base R 'hclust' function on the Euclidean distance between samples, drawn with ggdendrogram package (de Vries & Ripley, 2020). Samples highlighted in red were selected for exclusion from downstream analyses due to their dissimilarity to other samples from the same experiment based on the clustering and other signals.

Based on this rationale three samples were removed; a 2h Liver infection sample (SRR11142819) from Toro-Moreno et al., 2020, and two schizont samples

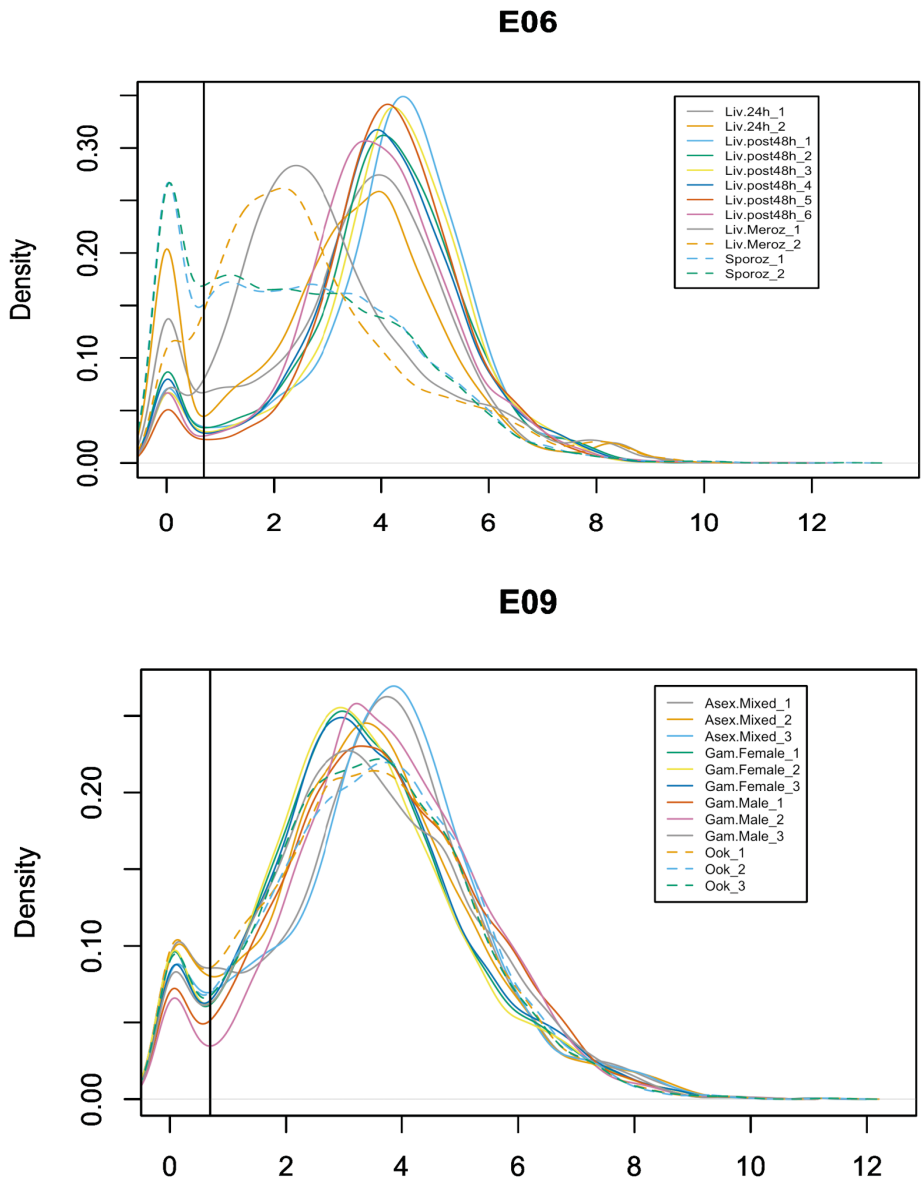


Figure 3.3 - Density plots of the RNAseq sample ln(TPM+1) values for E06 and E09. Experiment codes as in Table 3.1. Drawn using the average bandwidth of every sample from each experiment: The bandwidth used is 0.262, which determines the smoothing of the density line. The mean bandwidth and density profiles were calculated using the base R 'density' function. Vertical black line shows the value of TPM = 1 (ln2). The number of genes included in each distribution (N) is 5245, all the genes annotated in *P. berghei*.

(SRR3437888 and SRR3437912) from Modryzynska et al., 2017 (these hierarchical clusters are shown in Figure 3.2). Additionally, some of the different

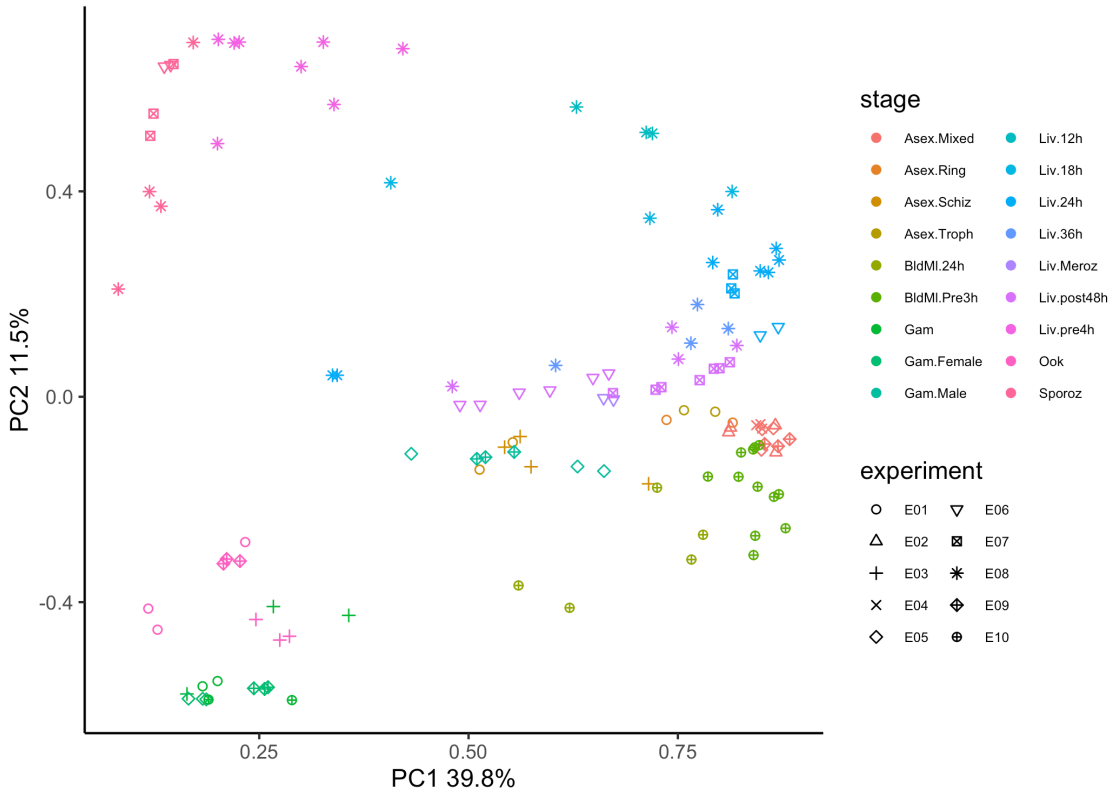


Figure 3.4 – Principal component analysis (PCA) plot of the genome TPM data coloured by the different stages and shapes determined by the originating experiment.
PCA was conducted using the 'prcomp' function in R with zero centring and scaling to unit variance. The x axis represents principle component 1 (PC1) while the y axis represents PC2. Percentages on the axes show the variance of the data covered by each component.

conditions/time points defined in the original publications were highly similar to one another. I combined the groups of these samples based on these similarities (See Materials and Methods). The remaining samples were then compared based on the distribution of TPM values per gene, as this is a metric that TPM does not control for (Hart et al., 2013). These distributions of gene TPMs for representative experiments (E06 and E09) are shown in Figure 3.3 as density plots. For the vast majority of life-cycle stages, the traces show a double peak; an initial peak of genes that are not transcribed above noise, and then a subsequent peak of genes being actively transcribed (Hebenstreit et al., 2014). The trough between both peaks could be seen as a measure of threshold transcription, above which a gene is being actively transcribed and below which it is indistinguishable from noise. For most experiments and samples,

the commonly used threshold of 1 TPM (the vertical lines in Figure 3.3) is a good estimator of this trough. Interestingly the sporozoite samples show distinct transcriptional traces compared to other life cycle stages, only clearly possessing a single peak at low TPM levels. Since this pattern is common to sporozoites across experiments, it is probably a biological result as opposed to an artefact. This is likely due to low numbers of transcripts in the sporozoite stage, with minimal active transcriptional activity and a relatively small repertoire of stored, repressed transcripts (Roch et al., 2004; Silvie et al., 2014).

After these initial quality control steps, the samples were analysed together to detect associations between published studies. Figure 3.4 shows the PCA (principal components analysis) plot of the transcriptional profiles of each sample from all experiments. From this plot, I saw how across 51.3% of the variation in the data that the samples associate closely with other stages from the same life-cycle stage. By constructing a biplot (see Figure 3.5 – this plots shows the PCA overlaid with the effect that each gene is having on the PCA results, so the location of each gene on the plot shows the direction it is influencing the samples in the principle component space), I observe that the distinct groups formed in the PCA were influenced by known marker genes, such as *chitinase* and *secreted ookinete adhesive protein (SOAP)* distinguishing the ookinete samples (Dessens et al., 2003; Huber et al., 1991), and *P25/28* transcription associating with female gametocyte samples (Paton et al., 1993). I then averaged the data across life-cycle stages, but within experiments. Figure 3.6 shows the correlation heatmap of these data, further confirming how transcriptional profiles from different stages of the *P. berghei* life cycle group together, and do not form strong associations with samples from the same experiment but different life cycle stages. This is especially clear for the distinct clusters of *P. berghei* life cycle stages, roughly described as: post 24h liver schizonts and asexual blood stage schizonts; male gametocytes; mixed asexual blood stages and 12-24h liver schizonts; sporozoites; ookinetes; and female/mixed gametocytes.

Most of the published studies of the transcriptomes of asexual blood stages of *P. berghei* use mixed samples, i.e. not separating the ring, trophozoite, schizont or merozoite development stages in this part of the life cycle. These already heterogenous samples could also contain the sexual stages of the parasite, which develop in the

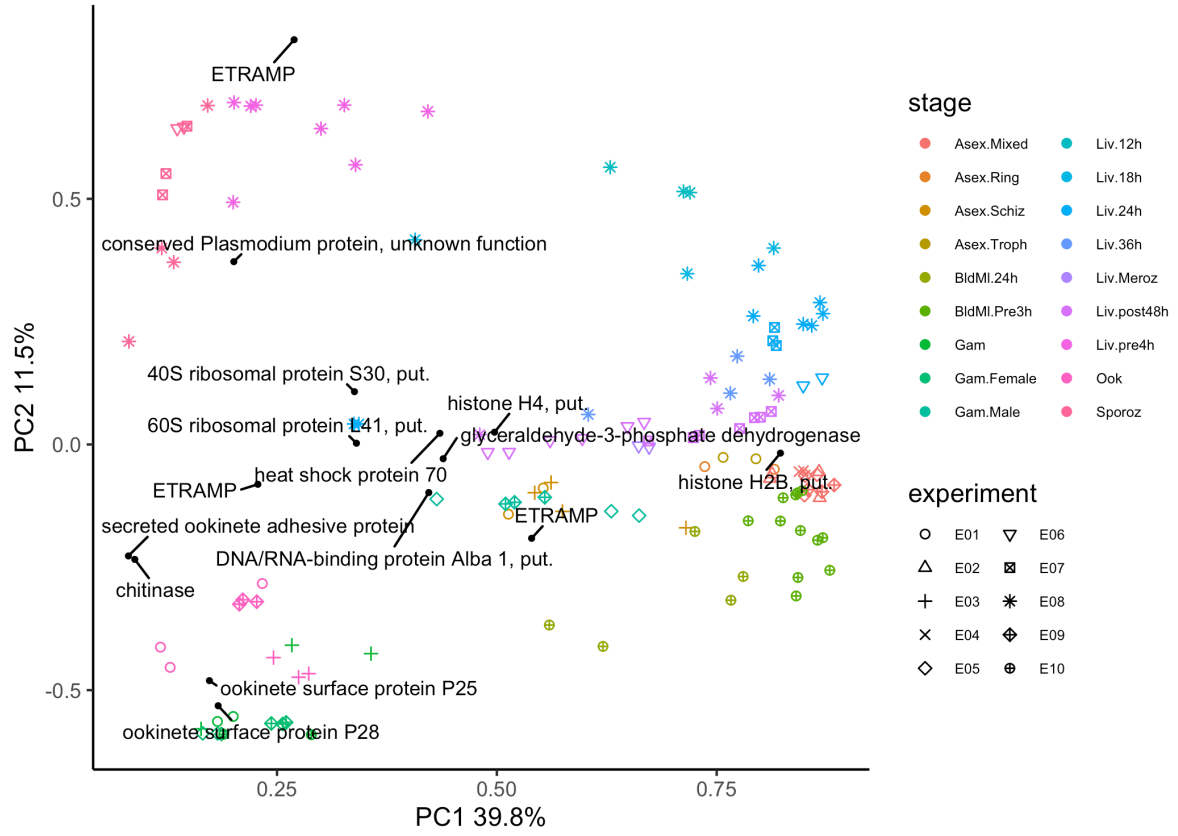


Figure 3.5 - Biplot of the *P. berghei* experimental samples, highlighting the top 15 genes influencing the PCA results.
PCA was drawn as in Figure 3.4 and then the loadings for the 15 genes with the highest absolute loading values were superimposed as points (the actual loading values were multiplied by a factor of 0.005 to fit onto the plot). Lines show which gene label corresponds to which loading. The position of the genes indicates in which 'direction' it 'pulls' the stages on the PCA. For example, the ookinete surface proteins in the bottom left of the plot indicate that these genes are transcribed highly in nearby samples and that this distinguishes them. We know from the sample IDs and the literature that this is characteristic of ookinetes and female gametocytes. *ETRAMP* stands for the *Early-Transcribed Membrane Protein family*. *P25* is the *25 kDa ookinete surface antigen precursor* and *P28* is the *28 kDa ookinete surface protein*. 'put.' Refers to a putative annotation.

circulation of the host. To ensure that the mixed asexual samples are not contaminated with sexual stages, I determined the transcription of four gametocyte markers (*p28* and *nek4* (*NIMA-related kinase*) – markers for female gametocytes and ookinetes

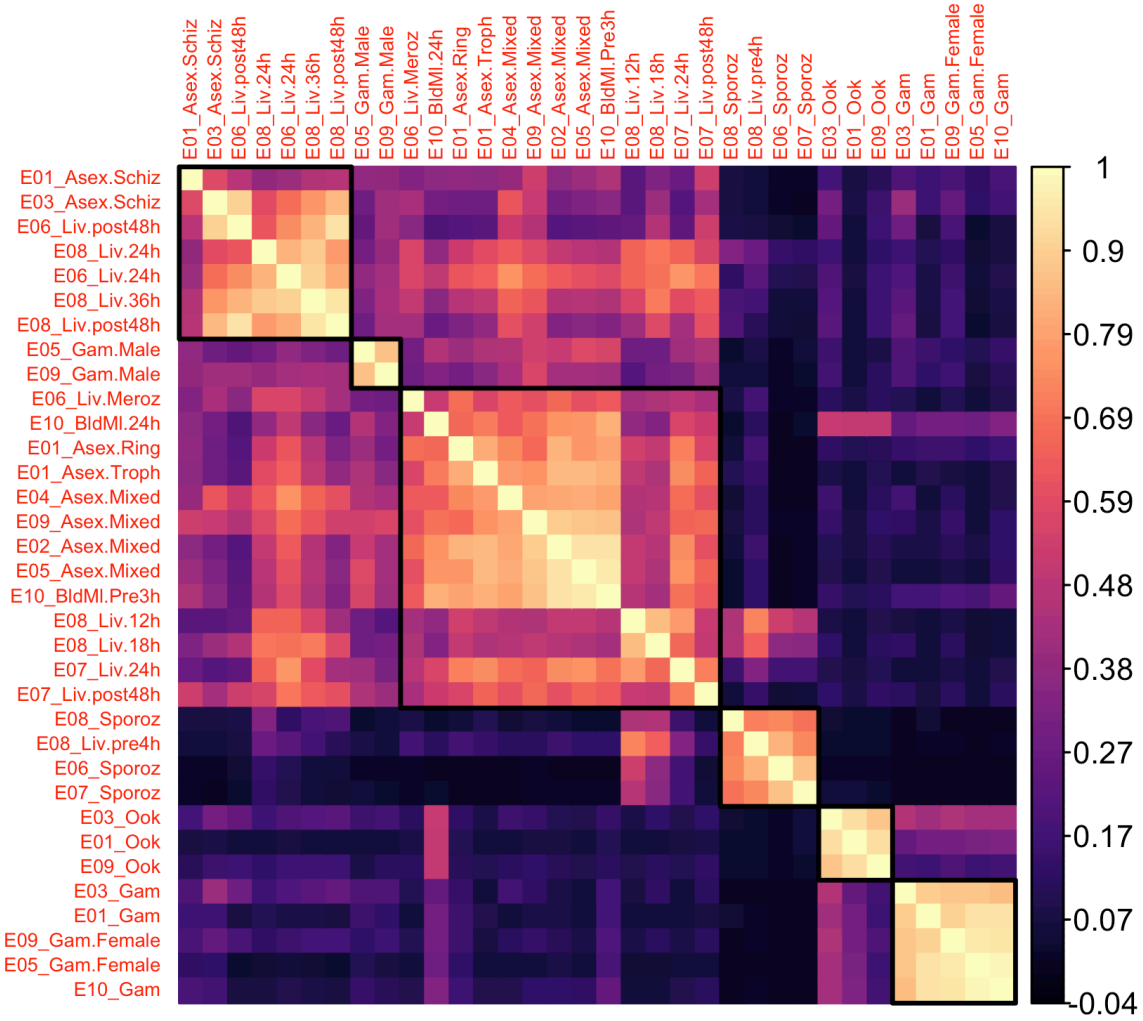


Figure 3.6 – Pearson correlation plot (-0.04-1.00) of each transcriptome calculated from genome TPM values for each stage from each experiment.
The colour of the boxes is determined by the Pearson correlation between every gene in each collated sample. The samples were clustered by hierarchical clustering using Ward's clustering criterion ('ward.d2' setting in corrplot function (Wei & Simko, 2017)), and six clusters are highlighted in boxes. Six clusters were chosen as most optimal with the 'elbow method' (Thorndike, 1953)**, as implemented through the factoextra package function 'fviz_nbclust' using k-means clustering and within cluster sums of squares.**

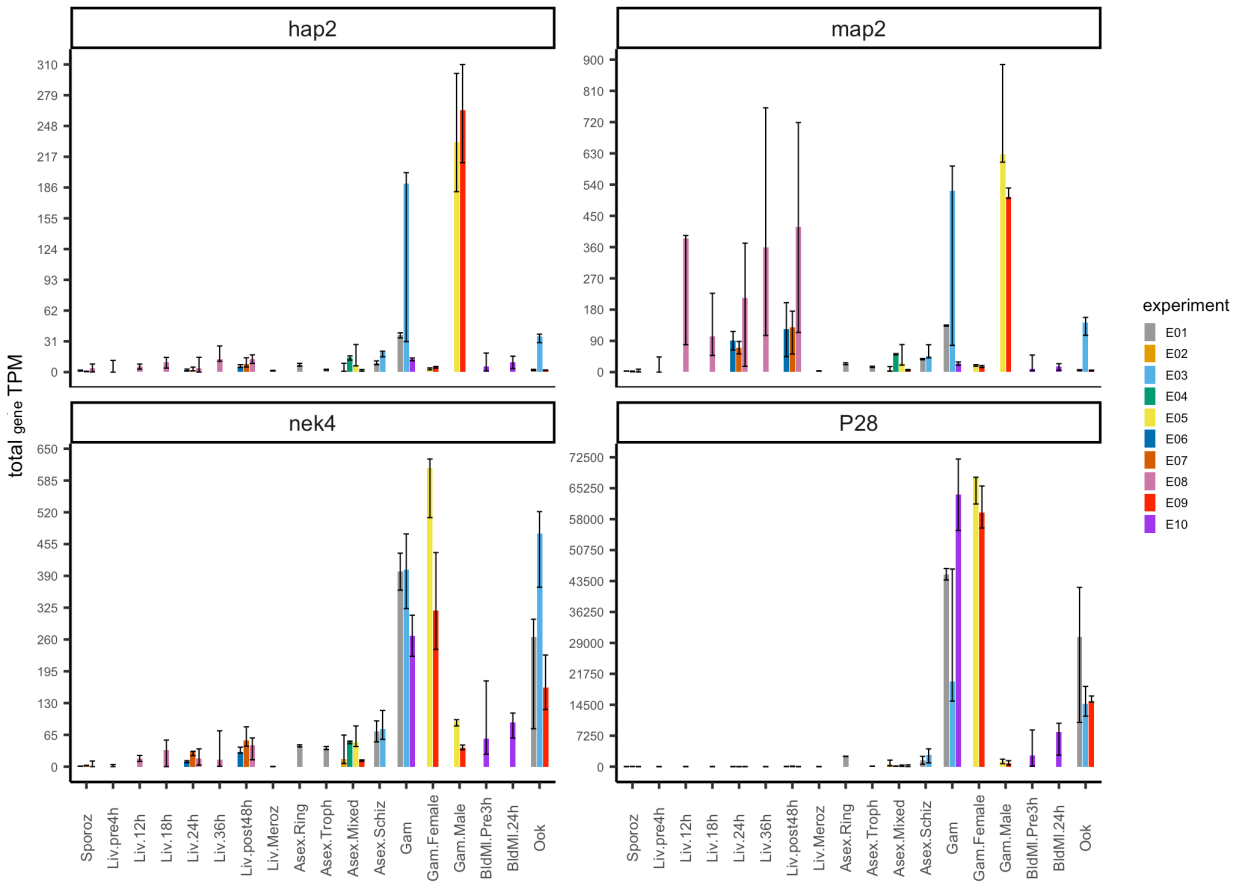


Figure 3.7 - Transcription of gametocyte markers among the *P. berghei* experiments and life cycle stages.

Transcription of four gametocyte-specific marker genes across the life cycle stages of P. berghei in the different experiments. The genes included are *p28* (*PBANKA_0514900*) and *nek4* (*NIMA-related kinase 4*) (*PBANKA_0616700*), markers for female gametocytes and ookinetes; *mapk2* (*mitogen-activated protein kinase 2*) (*PBANKA_0933700*), marker for male gametocytes; *hap2* (hapless orthologue) (*PBANKA_1212600*), gamete fusion protein. Bar height corresponds to median TPM, with error bars showing the range of TPM values across replicates. Colours correspond to different experiments. Abbreviations are the same as previous figures.

(Paton et al., 1993; Reininger et al., 2005); *mapk2* (*Mitogen-activated kinase 2*) – a marker for male gametocytes (Dorin et al., 1999); and *hap2* (hapless *orthologue*) – a marker for gametocyte stages in general (Liu et al., 2008). Figure 3.7 shows that the gametocyte markers are transcribed in the mixed asexual samples at only a low level, confirming the sample purity.

From the above analyses I conclude that the published experiments robustly captured the different stages of the *P. berghei* life cycle, and that the transcriptional profiles can be accurately compared between the studies. Additionally, the mixed asexual stages, though inherently heterogeneous, are pure captures of the asexual stages and have minimal contamination from gametocytes.

3.3.2 The total transcription of *pir*s across the life cycle of *P. berghei* consistently oscillates with given sub-family enrichment at certain stages

The broadest method by which to analyse the transcription of a multigene family is to look at the family as a whole. I used this to analyse when the *pir* genes are most needed across the *P. berghei* life cycle, and when they likely are not required. From this overview I then used smaller sub-classifications of the family to investigate how their transcription is altered by time of development, finding patterns which indicate when given subsets of *pir*s are transcribed. Finally, I asked how individual *pir* genes change transcription across the life cycle.

Overall, the *pir* gene family is transcribed at a lower level than the rest of the genome in every stage of the life cycle (Figure 3.8). The median TPM is lower than that of 134 randomly selected genes from across the transcriptome. A comparison of transcription of the families of *P. berghei pir* genes shows that the sporozoites and early liver stages transcribe almost no *pir* genes, however transcription of *pir* increases with progression through liver stage schizogony (Figure 3.9). There is a high degree of variation in the experimental values of the *pir* TPM at these early stages, suggesting that they sometimes transcribe *pir* genes (high maximum) but usually do not (low median). This increase in overall transcription is concomitant with an increase in the number of distinct *pir*s being detected at an ‘active’ level. An upsurge of *pir* transcription is observed in the liver merozoites that burst out of the liver schizonts (note that this is from one experiment (Caldelari et al., 2019) and hence there are no inter-study replicates). This represents the first major peak of *pir* transcription in the *P. berghei* life cycle. A high level of *pir* transcription is maintained throughout the asexual stages, with the separated asexual blood stage data (with the ring and trophozoite stages again from a single RNAseq publication (Otto et al., 2014)) suggesting that

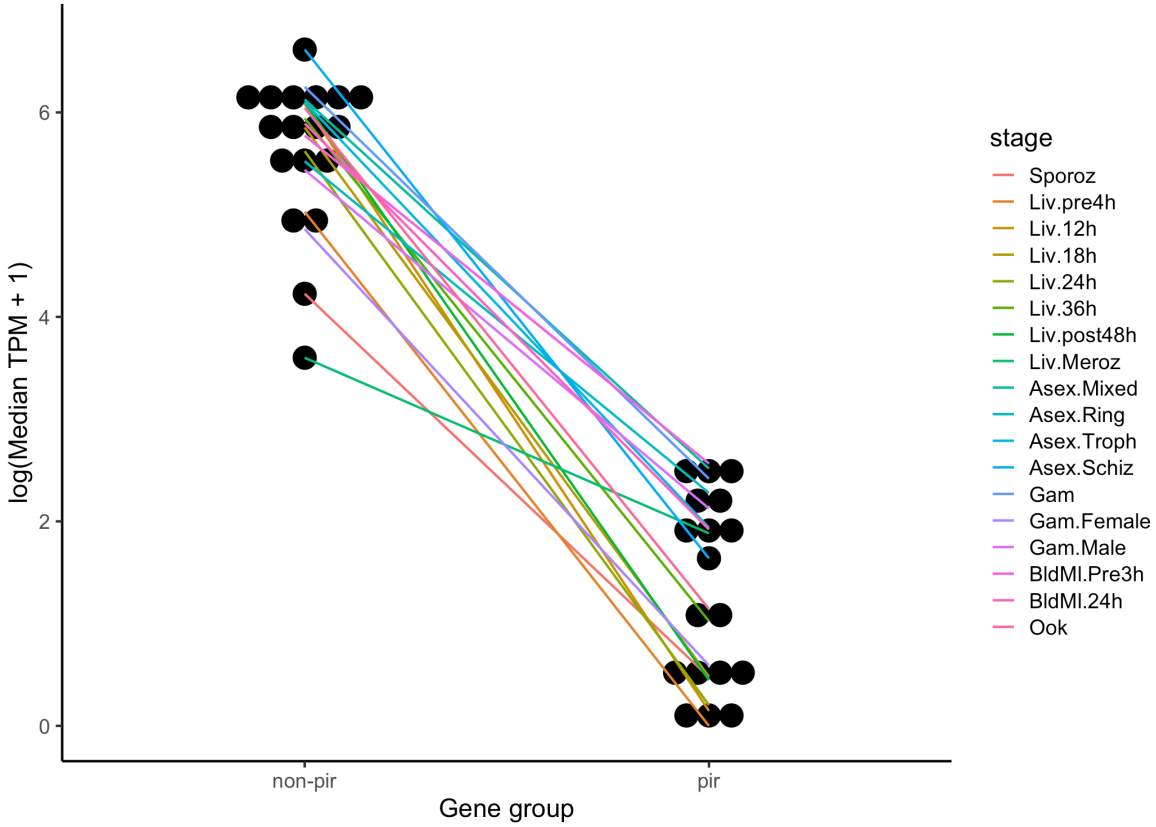


Figure 3.8 - The median TPM of *pir* genes (right) and 134 randomly selected non-*pir* genes (left) across the life cycle.
Each dot represents the median TPM for a given life-cycle stage across either the random non-*pir* genes or the *pir* genes. Lines link median values from the same stage of the life cycle, coloured as shown in the legend. Life-cycle stage abbreviations include: Sporoz (Sporozoite); Liv (Liver stage); Liv.Meroz (Liver stage-released merozoites); Asex (Asexual blood stages); Gam (Gametocytes); BldMl (Mosquito blood meal); Ook (Ookinetes).

there is a decrease of *pir* transcription from ring to schizont stages. At the sexual stages of the life cycle, a stark difference can be observed between the different sexes of gametocytes, as observed by Reid and colleagues, 2018. There is a diminished transcription of *pir* genes in the female gametocytes, while the male gametocytes represent another peak of *pir* transcription. The standard deviations of the female gametocyte TPM values indicate that variation is quite low.

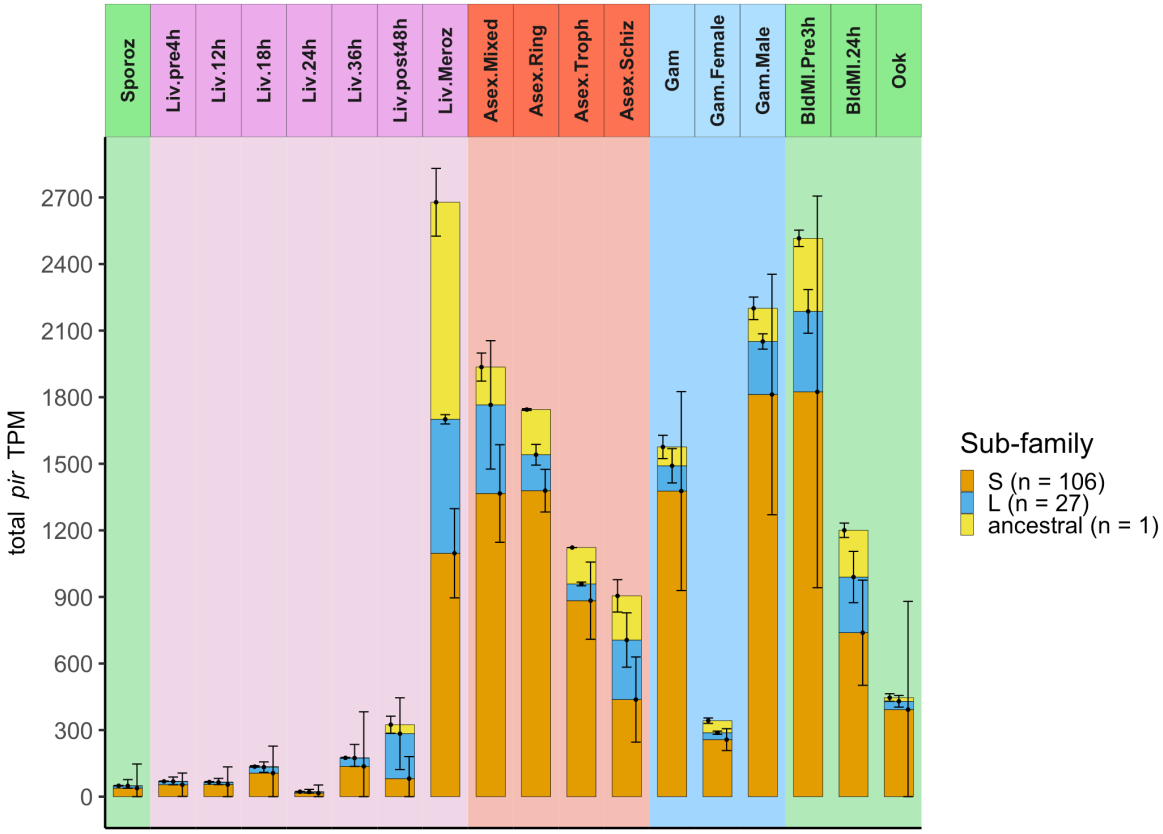


Figure 3.9 - Transcription of *pir* genes throughout the *P. berghei* life cycle, split by sub-family designation.
Error bars show one standard deviation across all samples (restricted so they cannot go below zero). The y-axis shows the mean *pir* transcription for each sub-family as a stacked bar chart. n = the number of genes in each sub-family.

This suggests that the few *pir*s being transcribed are likely to be of importance, as female gametocytes are always found with at least some *pir* transcripts. At the mosquito stages of development (the ookinetes) I see relatively low levels of *pir* transcription, similar to the female gametocytes and early liver stages. The ookinete *pir* transcriptome is actually skewed to be higher by the results of one experiment (Modrzynska et al., 2017) which I believe is more gametocyte-like in profile (suggested by the presence of gametocyte-exclusive markers in Figure 3.7), perhaps due to immature ookinetes or gametocyte contamination. This is shown by the high variation of the Short-form *pir* transcription at this stage, which is due to a mix of samples with high and low transcription. Overall, both the mosquito stages, the ookinetes and

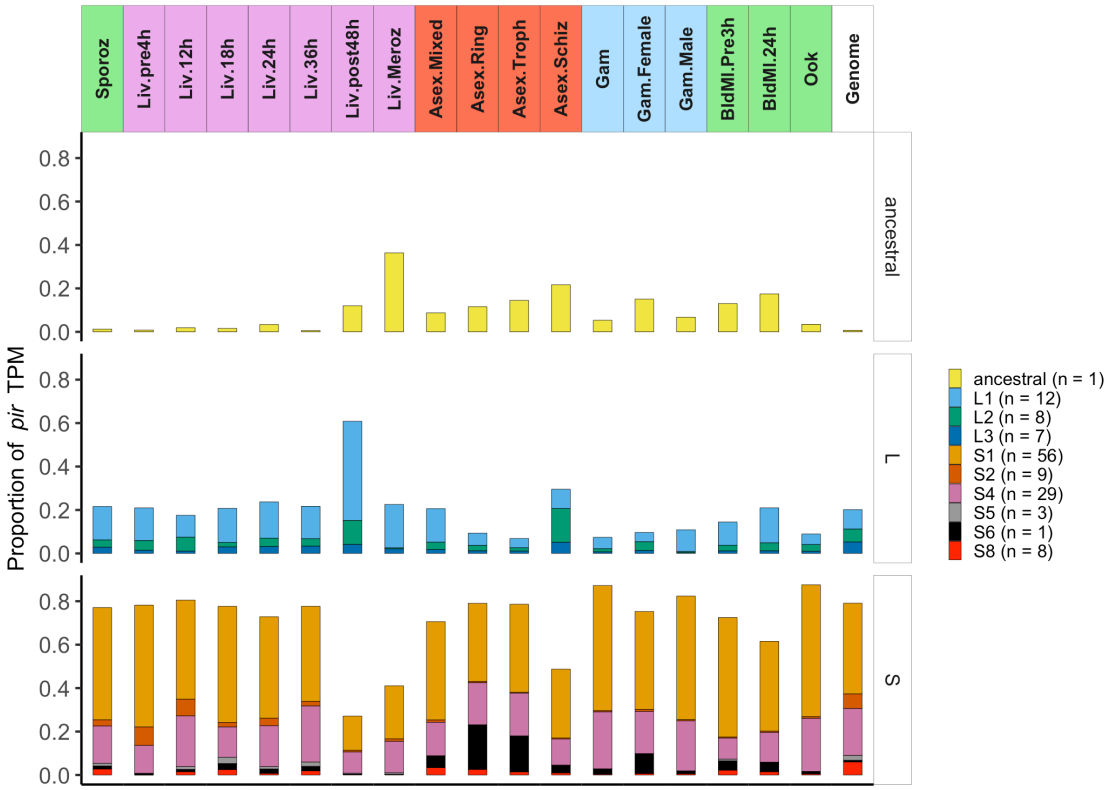


Figure 3.10 - Proportions of total *pir* TPM contributed by different sub-families across the life cycle.
The final column represents the composition of the genome. Note that each column (life cycle stage) across the three sub-families adds up to one, as this is showing proportional data of the total. N = the number of genes in each sub-family.

sporozoites, show minimal *pir* transcription, which has been observed before (Otto et al., 2014).

Three major peaks of *pir* transcription are suggested by these data; liver merozoites, asexual blood stages (which overlap with the early mosquito blood meal), and male gametocytes. Figure 3.9 and Figure 3.10 also show how the transcripts are partitioned into different *pir* sub-families; the Short (S), Long (L) and ancestral (a single *pir* believed to be the ancestral sequence) (Otto et al., 2014). The proportions of these different families in the RNAseq profile vary somewhat between different stages of the life cycle, such as 7-9% of L-form transcription in the rings and trophozoites (from Otto et al., 2014) increasing to 33-34% in schizont samples (although the change of expression is not statistically significant – see below). One of the starkest changes in sub-family proportion between life-cycle stages can be observed in the post-48h liver stages, which

is enriched for L-form *pir*s (61%). I can further sub-classify the *pir*s to find a greater granularity in these observations (see Figure 3.10). The post-48h liver stage enrichment of L-forms is mostly due to the L1 group of Long *pir*s. This contrasts with the asexual schizonts in which the higher L-form transcription is dominated by the L2 group. The S-forms are transcribed to the greatest extent in most stages of development, especially the S1 group which is the largest sub-family group in the genome.

Differential expression analysis was then employed to find the changes in sub-family transcription between individual life cycle stages which were supported statistically. These comparisons of groups of genes require large enough sample sizes of gene sets, so the sub-families with fewer than 5 members (ancestral, S5 and S6) were not included. The comparisons chosen (Figure 3.11) were based on consecutive stages of the life cycle, but I also directly compared male and female gametocytes, as well as the major peaks of transcription (Gam.Male, Asex.Mixed and Liv.Meroz). As demonstrated in Figure 3.11 there are several differentially expressed *pir* sub-families between the life cycle stages, however most of these changes follow the overall *pir* transcriptome upregulation or downregulation. For example, multiple *pir* sub-families, and the *pir* genes overall, are significantly upregulated in the liver merozoites (Liv.Meroz) compared to late liver schizonts (Liv.post48h), as well as in the male gametocytes (Gam.Male) compared to female gametocytes (Gam.Female), which is expected from Figure 3.9. The enrichment of the Long-form *pir*s in the late liver schizonts (Liv.post48h) is confirmed by this analysis. Although there are no significant differences between gene set enrichment of the 36h and post-48h samples, by comparing post-48h liver stages and 24h liver stages I see that there is a much greater log-fold increase of L-form *pir* expression than for the *pir*s overall (all *pir*s between liver post 48h and 24h: log2 fold change 0.67, FDR 0.004; L-form *pir*s: log2 fold change 1.56, FDR < $1x10^{-10}$; S-form *pir*s: non-significant). Additionally, the only *pir* sub-family not significantly upregulated in liver merozoites compared to mature liver schizonts is the L2 sub-family (with overall a slightly negative log fold change). Differential expression results that defy

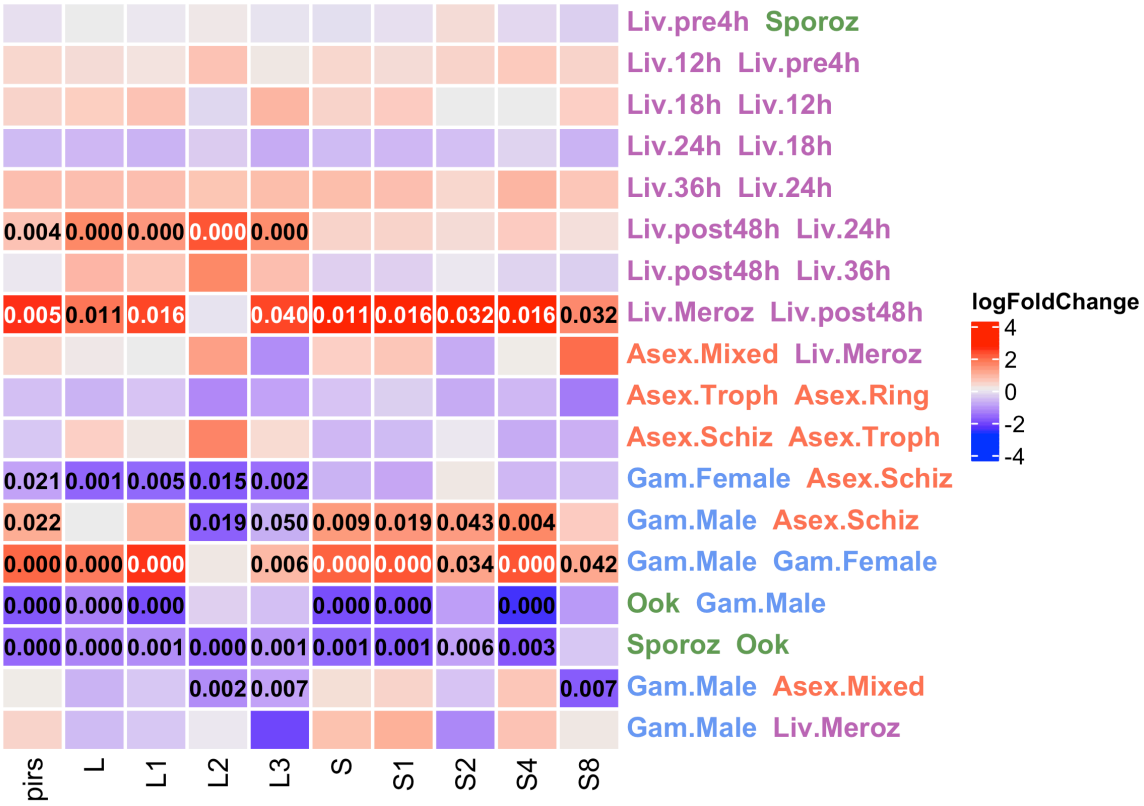


Figure 3.11 - Heatmap of the log-fold change of the *pir*s and *pir* sub-families between sequential and selected stages of the *P. berghei* life cycle.
Numbers within the boxes show the adjusted p value of statistically significant differential expression comparisons (adjusted p < 0.05). The comparison is between the first and second stage in the names (on the right-hand side of the heatmap), so for example the positive log-fold changes between Gam.Male and Liv.Meroz show upregulation in the Gam.Male compared to Liv.Meroz, and vice-versa.

the overall change of the *pir* gene family are the L2 and L3 groups between asexual schizonts and male gametocytes (Asex.Schiz and Gam.Male). There is a general upregulation of *pir*s if asexual schizonts differentiate into male gametocytes (assuming that asexual schizont profiles are similar between sexually-committed and uncommitted cells), however the L2 *pir*s are significantly downregulated in the male gametocytes compared with asexual schizonts. Between the major peaks of *pir* transcription, there are only statistically significant changes between the male gametocytes and asexual stages: lower transcription of L2, L3 and S8 *pir*s in Gam.Male vs Asex.Mixed. The S8 *pir*s are also expressed to a much lower extent in the liver merozoites compared to

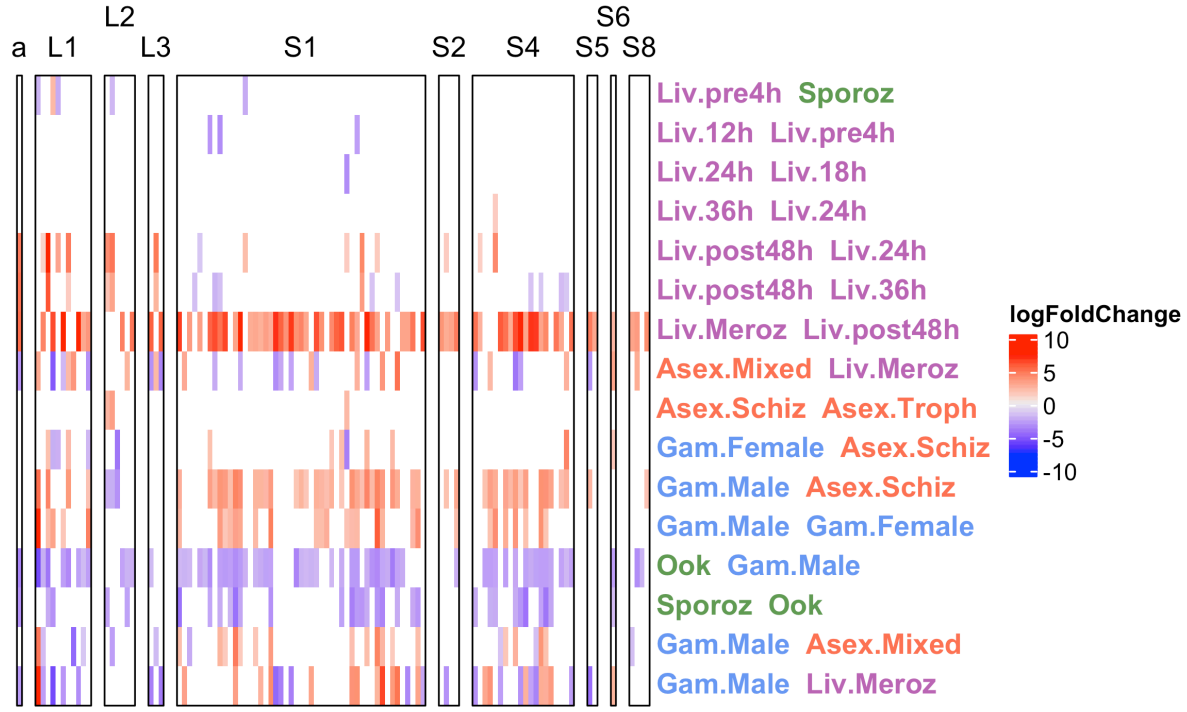


Figure 3.12 - Heatmap of the log fold changes of individual *pir* genes with statistical evidence for differential expression (FDR < 0.05) between given stages of the *P. berghei* life cycle.

The log fold change shows the difference in the former stage compared to the latter, so a positive log fold change (in red) of a *pir* in the Liv.Meroz_Liv.post48h row shows that the *pir* is significantly upregulated in liver merozoites compared to late liver schizonts. The *pir*s (each column) are split into the sub-families (any sub-family not included have no members with significant differential expression). The 'a' sub-family refers to the 'ancestral' *pir PBANKA_0100500*.

mixed asexual stages (Asex.Mixed), however this, and other liver merozoite comparisons, do not reach significance because of the low sample size of liver merozoites (n=3). So, it cannot be concluded from these data whether samples of asexual mixed-stage parasites are indeed enriched for S8 *pir*s, but it is strongly suggested by the expression evidence. Overall while there are some specific enrichments for sub-families in given stages of the life cycle, they are mostly either upregulated or downregulated together, with few occasions when there is statistical evidence for both up and down regulation of sub-families in the same stage transition. These *P. berghei* data sets provide evidence that there is specific enrichment for the expression of Long-form *pir*s in late liver schizonts, S8 *pir*s in mixed asexual stages, and L2 *pir*s in asexual blood schizonts.

When performing differential expression of individual genes, a similar pattern is observed, whereby most of the genes with altered transcription across a given life-cycle transition are either all upregulated or downregulated (Figure 3.12). The conversion from late liver schizonts to liver merozoites remains the most dramatic; a huge upregulation of *pir* gene expression is induced. Notably, most of the Long-form *pir*s which were upregulated as 24h/36h liver stages mature into Liv.post48h stages, are not further upregulated as the Liv.post48h stages differentiate into the liver merozoites. This may indicate that the general enrichment observed of Long-form *pir* RNA in late-liver schizonts is caused by a subset of the genes. Two of these L2 genes are also among the three L2 genes that are significantly downregulated in male gametocytes compared to asexual blood schizonts. Long-form *pir*s may be among the earliest *pir*s to play a role in mammalian invasion, as there is differential expression here even between early liver stage (Liv.pre4h) and Sporozoite samples (two L1s, an L2, and an S1 downregulated; one L1 upregulated).

Some of the sub-family comparisons which were indicated to be significant in Figure 3.11 give conflicting results when the differential expression of individual genes was examined. Expression of S8 *pirs* was found as a group to be significantly upregulated in mixed asexual stages and male gametocytes; however only a single individual gene is identified. This may be a consequence of enhanced power from combining the gene set; whereby individual genes do not change enough to give sufficient evidence for differential expression, but their combined signal is enough to conclude that the gene subset is differentially expressed. Sometimes members of the sub-families are changing transcription in different directions, even when the gene-set statistics showed an overall trend one way. An example is the significantly downregulated expression of L1s between female gametocytes and asexual schizonts, shown in Figure 3.12, which turns out to be a mix of both significantly up- and down-regulated genes. The transcriptional changes of members of sub-families are not always correlated with one another. This can be especially seen in the comparisons between the highest *pir* expressors, where different members of the same sub-families may be either up- or down-regulated as the parasite progresses through the life cycle.

Certain individual *pir*s showed a strong association with given stages of the life cycle. As examples, the L1 *pir PBANKA_0317181* was significantly upregulated in male gametocytes (compared to mixed asexual stages: log fold change 4.95, FDR < $5x10^{-9}$;

compared to liver merozoites: log fold change 7.52, FDR < $1x10^{-6}$), and the L1 *pir* *PBANKA_0600031* showed enrichment in the liver merozoite samples (compared to mixed asexual stages: log fold change 5.25, FDR < $5x10^{-6}$; compared to male gametocytes: log fold change 5.72, FDR < $5x10^{-5}$).

The *pir* genes experience major changes in transcription across the life cycle, consistently showing peak transcription during the liver merozoite, asexual blood and male gametocyte stages. While there is an upregulation of *pir*s as liver schizonts develop, concentrated in the Long-forms, there is a dramatic upregulation of almost the entire family upon the burst of the merosomes containing the merozoite forms which initiate blood-stage infections. Female gametocytes are particularly distinct from the males in their *pir* repertoire, as there is a major dip in *pir* transcription upon the differentiation into female gametocytes. Overall, the strongest signals of differential expression occur when large swathes of the family are being up or down-regulated collectively. Some patterns of enrichment, such as S8-enrichment in the mixed asexual stages, may be obscured by low statistical power.

3.3.3 The ancestral *pir* gene is consistently transcribed at high levels across the *Plasmodium* genus.

The above sections focussed on the major *P. berghei pir* sub-families; the Long and Short *pir*s. The single 'ancestral' *pir* (*PBANKA_0100500*) is kept in a grouping of its own as it is phylogenetically distinct from its genetic compatriots, and indeed its transcription stands out from the rest of the family. From both Figure 3.9 and Figure 3.10 it is clear how highly transcribed the ancestral *pir* is when *pir*s are generally upregulated. This individual *pir* is the most transcribed *pir* in most life-cycle stages, and individual samples. The ancestral *pir* has been highlighted before because of its high degree of conservation across *Plasmodium sp.* (Frech & Chen, 2013). The fact that an (often syntenic) ortholog of this *pir* is found across almost all genomes which contain canonical members of the family demonstrates its claim to being the parental gene from which the *pir* lineage initially emerged. Expanding upon earlier work, a particular region of conservation was aligned between the ancestral *pir* ortholog of multiple *Plasmodium* genomes including those only recently sequenced (Figure 3.13). The alignments contained numerous sections of high (90-100%) conservation, which is unusual for this variable gene family.

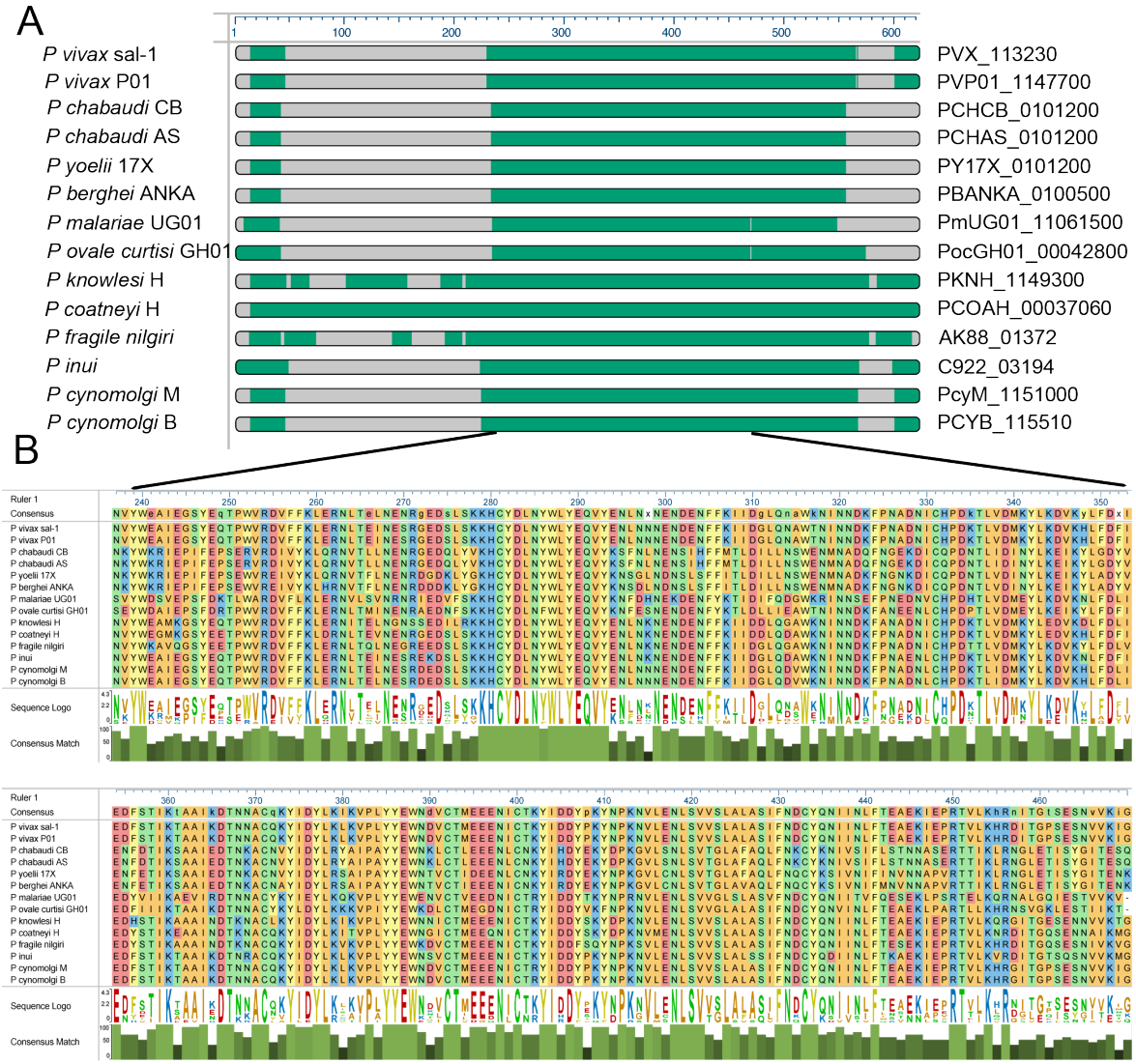


Figure 3.13 - Alignment of the ancestral *pir* from multiple *Plasmodium spp.* shows high degrees of conservation.

A. 15 orthologs of the ancestral gene from *P. berghei*, *PBANKA_0100500*, were included and aligned to show regions of homology. B. Close-up of the 234 aa block which is the most highly conserved of them all. A sequence logo and percentage match of the consensus sequence is included below the sequences. Data analysed and figure created by Deirdre Cunningham in (Little et al., 2021)**.**

As this high degree of transcription of the ancestral *pir* is also a feature of the *P. chabaudi chabaudi* AS transcriptome across the asexual cycle (see Chapter 4), I

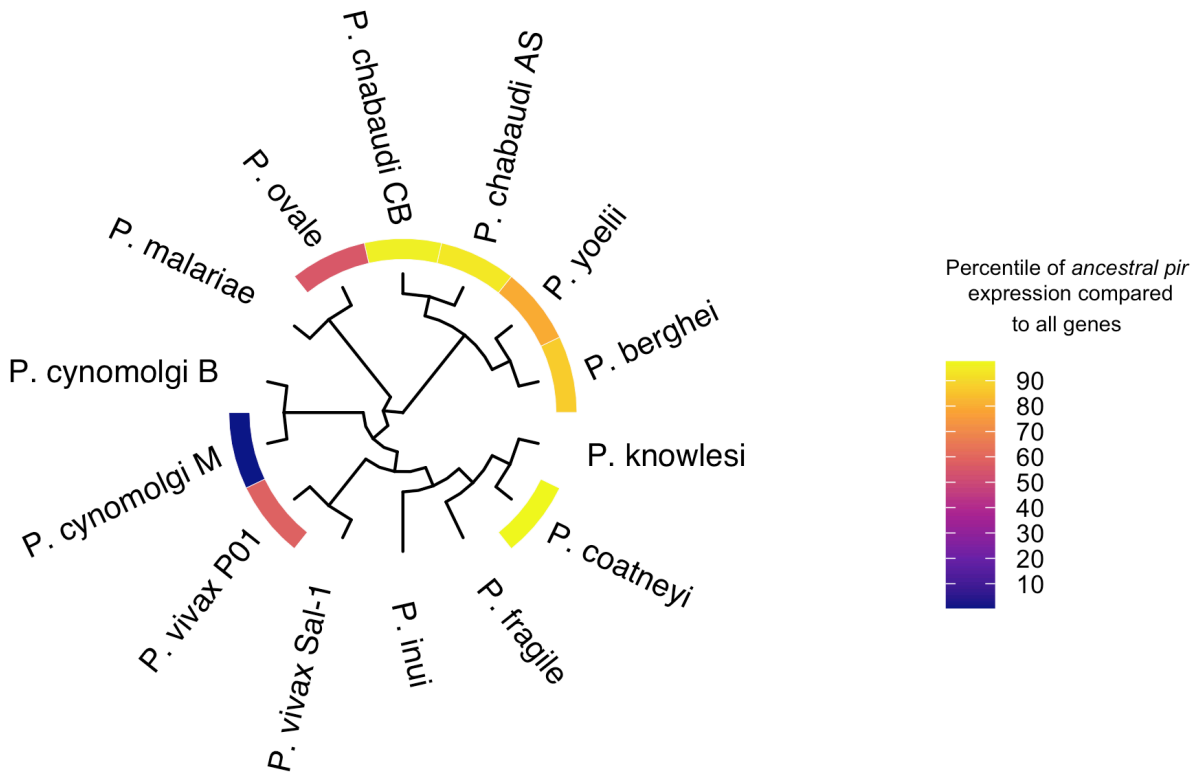


Figure 3.14 - Phylogenetic and gene expression analysis of the ancestral pir gene. Cladogram of the 15 orthologs, with heatmap of expression of the gene relative to the rest of the genome, where data is available (from multiple transcriptomic datasets as RPKM (reads per kilobase of transcript) or TPM – see Table 3.2).

hypothesised that the ancestral *pir* ortholog in each species would be among the most transcribed genes in RNAseq data. To investigate this, the RNAseq datasets of species with published data were collected and processed to calculate whether the ancestral *pir* was transcribed among the higher transcribed genes in the whole transcriptome (see Table 3.2). Where possible, data from asexual blood stages were included, however they were not always available. While this comparison uses both TPM and RPKM as input (depending on the results included with the original paper), I believe that this is sufficient for comparison within transcriptomes. This analysis shows that for most species the ancestral *pir* is detected within the top 50% of all gene transcription, often higher than 70%. The heatmap of transcriptome percentile is included alongside a phylogeny of the ortholog to show how it matches the lineages expected given the tropism of the species (separating by simian and rodent malarias). Species with lower relative transcription of the ancestral *pir* include *P. vivax*, *P. ovale* and *P. yoelii*, which notably have thousands of *pir*s in their genome, suggesting that there may be enhanced

Reference	Species	Stage
(Siegel et al., 2020)	*P. vivax*	Asexual.Schizont
(Hart et al., 2019)	*P. yoelii*	Gametocytes
(Ramaprasad et al., 2019)	*P. vinckei*	Asexual.Mixed
(Lin et al., 2018)	*P. c. chabaudi* CB	Asexual.Mixed
(Brashear et al., 2019)	*P. ovale*	Asexual.Mixed
(Voorberg-van der Wel et al., 2017)	*P. cynomolgi*	Liver.Schizonts/Hypnozoites
(Cordy et al., 2019)	*P. coatneyi*	Asexual.Mixed

Table 3.2 - Published studies from which transcriptome data was taken from different *Plasmodium spp.* and used in Figure 3.14.

redundancy of function of the ancestral *pir*. The lowest transcription of the ancestral *pir* relative to total transcription is in the *P. cynomolgi* RNAseq data, however these RNAseq data are obtained from liver stages in which I would expect (based on the *P. berghei* data) that transcription could be lower.

3.4 Conclusions

By using multiple RNAseq datasets collected from different published studies, I was able to analyse the dynamics of the *pir* transcriptome transcending most of the life cycle of *P. berghei*. With this collection, I observe that the peak of *P. berghei pir* transcription occurs during the liver merozoites/merosomes, the asexual stages, and in male gametocytes. There are troughs of *pir* transcription during the mosquito stages, with sporozoites transcribing virtually no *pir*. These differences are observed consistently across different experiments. Sub-family changes are associated with certain stages, such as enhanced Long-form transcription in liver and asexual schizont development (corroborating the results of Otto et al., 2014 for the latter), especially mediated by L2s, while Short-forms such as the S8s are possibly enriched in asexual blood stages. This may attest to different functions of these cladistic groups. An outlier to the generally low *pir* transcription was also the phylogenetic outlier, the ancestral *pir*, which shows high transcription when *pir*s are generally transcribed. The syntenic orthologs of this conserved *pir* across different *Plasmodium spp.* also show high transcription.

A study such as this could be limited by batch effects between published studies, however clustering analysis performed here demonstrated that the batch effects between experiments were not strong enough to overcome the biological signals of different *P. berghei* life-cycle stages. I chose to use this as evidence that no more correction or normalisation was required, however there are methods available which may be able to remove any batch effects present and provide additional robustness. Trimmed Means of Medians (TMM) is one such method, which corrects for the distributions of transcripts in samples – which would also be useful for correction between samples even within the same study (Robinson & Oshlack, 2010). This method is predicated on the assumption that most genes are not differentially regulated, which is an assumption I believe does not hold for different life-cycle stages of *Plasmodium spp*. Other examples include ComBat-seq, which uses methods from edgeR to estimate the mean batch effect and batch dispersion, and then maps the counts to a negative binomial distribution, which often fits count/TPM data (Robinson et al., 2009; Y. Zhang et al., 2020), and Remove Unwanted Variation (RUV), which uses control genes/samples which are not expected to vary between the experimental conditions for correction (Risso et al., 2014). It is impossible to fully account for variation between different experimental results because many of the studies do not cover overlapping life cycle stages with other experiments for direct comparison. Nonetheless, for *P. berghei* I was able to verify that the major variance in the samples is due to the different life cycle stages, and not due to experiment.

Many of the techniques used to obtain sufficient material for RNAseq from certain life cycle stages of *P. berghei* will have an unknown impact on the transcriptome of the parasites. All the gametocyte samples, and the blood used for ookinete samples, were taken from phenylhydrazine-treated mice (in order to stimulate reticulocyte production and gametocyte formation), except for Otto *et al*., 2014, which did not report using phenylhydrazine. One can only account for so much when doing these extractions, but it needs to be mentioned that this is an additional variable compared to other blood stages. Modrzynska et al., 2017 used phenylhydrazine for the schizont samples as well, which could offer an opportunity to investigate the impact of phenylhydrazine directly on the parasite transcriptome independently of development. There may also be differences induced between samples as a consequence of *in vitro* culture. Brief culture was required for extraction of synchronised asexual blood cycle stages, since the *P. berghei*

circulatory infection is naturally asynchronous. Ookinetes too were generated not *in vivo* but by inducing gametocytes to begin sexual reproduction in 'ookinete media'. The culturing of the liver stages uses a host considerably distinct from the natural parasite environment. All experimental datasets in this thesis were from infecting human-derived cell cultures with sporozoites, such as HeLa cells, and not rodent cells. The exoerythrocytic forms of the parasite are remarkably versatile, being able to grow not just in human cells but can be induced to form *ex situ* without being inside a cell at all (Kaiser et al., 2003). Evidence from *P. yoelii* suggests that parasites develop normally inside human hepatoma cell culture (this species does not grow in HeLa cells), indicated by their ability to form infectious merozoites (Calvocalle et al., 1994). I am not aware of infectivity being tested for *P. berghei* merozoites grown in human tissue. Primary liver culture from mice is a laborious and time-sensitive process, so perhaps immortalised human liver cell lines (such as hepatomas) or immortalised mice liver cell lines (Risal et al., 2011) would provide closer physiological matches to the natural environment of *P. berghei* exoerythrocytic stages. Single-cell RNAseq taken from *in vivo* conditions may also be able to help elucidate whether the methods required to extract enough material for bulk RNAseq impact the parasite transcriptome directly.

The transcription of *pir*s is much higher in the rodent blood than during the mosquito stages of development, or during exoerythrocytic growth. The lower transcription of *pir*s in the mosquito compared to the vertebrate stages has been reported before (Otto et al., 2014; Witmer et al., 2020), and this body of evidence continues to suggest that *pir* genes are not transcribed in the insect vector. Since *pir* transcription is low during this entire section of the life cycle (as far as they have been covered by transcriptome studies) it is likely that the *pir* genes do not have a constitutively-required housekeeping function, or any function that is required for propagating within the vector, such as motility. There is the possibility, however, that these data available for *P. berghei* may be hiding a role of PIR proteins at this stage. *Plasmodium* parasites frequently repress translation such that the transcription of a gene may be temporally decoupled from production of the protein (Mair et al., 2006a). The high transcription of *pir*s in the ookinete's predecessor, the male gametocyte, may include transcripts which are translated later in the life cycle. Only studies looking at the protein level would be able to discern the presence of PIRs at this stage. The earliest proteomic study found multiple PIR peptides in *P. berghei* ookinete, oocyst and sporozoite results (Hall et al., 2005), and PIRs have also been found in more recent studies of *P. berghei* sporozoite stages (Hamada et al., 2021).

Given the negligible transcription of *pir* genes in the RNAseq samples included here, the appearance of any PIR in the sporozoites is unexpected. Indeed, proteomic analysis of both oocyst and salivary-gland sporozoites of *P. yoelii* (which has about a thousand *pir* genes in the genome) detected no PIR peptides at all (Lindner et al., 2019). However, since overall transcription is so low in the sporozoites, it is possible that the mRNA production for any *pir*s translated occurs earlier. This has been observed for the single *P. falciparum* Erythrocyte membrane -1 (*Pf*EMP1) protein found on *P. falciparum* sporozoites, which is essential for liver cell invasion but not actually detected to be transcribed in the sporozoite stage. Instead, the sporozoite *var* gene is transcribed during the late oocyst stages, when sporozoites are formed (Zanghì et al., 2018). In fact, there is some evidence that the oocysts may be a focus of *pir* transcription during the *P. berghei* life cycle, as seen in the earliest oocyst stages analysed by the single-cell malaria cell atlas (Howick et al., 2019). This stage of the life cycle is absent from all bulk RNAseq studies, a knowledge gap which should be rectified by investigating a time course post-ookinete.

Although the liver schizonts are predominantly only transcribing low levels of *pirs*, there is a great upsurge of Long-form *pir*s in the later times of liver-stage development. There are no studies of any Long-form PIR protein expression in liver samples, so it remains to be confirmed experimentally to see if the PIRs are indeed being translated. IFA studies of 30/44/48h timepoints of liver stages of *P. berghei* found evidence of the presence of 1 out of the 4 canonical Short-form PIR proteins investigated, an S4 PBANKA_1400300 (another PIR identified as transcribed during the liver stages, PBANKA_0524600, is not a canonical PIR, as discussed below. (Fougère et al., 2016)). The same study found 5/5 PIR proteins in the asexual blood stages (which matches the high transcription at this point in time) and the singular PIR investigated in oocyst and sporozoite stages to be absent. There is a huge peak of transcription of most of the *pir* gene family once merozoites burst forth from the mature liver schizonts. Hence it may be expected to see multiple PIR proteins in the liver merozoite proteome, but despite this, there is only peptide evidence for the top expressed *pir* – the ancestral protein (Shears et al., 2019). This may be a consequence of the low sensitivity of current proteomic methods, which often only uncover an upper proportion of the proteome. It is known that the *pir* genes tend to be under-transcribed compared to the rest of the transcriptome, so it is possible that the proteins are only found at relatively low levels too. Therefore, the highest

transcribed *pir*, the ancestral *PBANKA_0100500*, would be the one most likely to appear in proteomic studies.

The difference in *pir* transcription between the female and male gametocytes presents the tantalising possibility that the PIR proteins may play a role in fertilisation. This conclusion from these bulk RNAseq data supports similar observations from single-cell data (Reid et al., 2018), but with the higher coverage afforded by the bulk method. However, from differential expression analysis between the male gametocyte stages and other high *pir* expressors, it seems that the peaks of *pir* transcription express many of the same sub-families and *pir* genes. This suggests that few of the *pir*s upregulated in male gametocytes could have a role only in fertilisation. I did find an individual male gametocyte-enriched *pir PBANKA_0317181*, which is a candidate for having a more specific function in male gametocytes.

The differential expression comparisons used above include many pair-wise tests. Adjusting for multiple testing is an essential part of the statistical analysis in order to have confidence in the results. However, this presents a problem for the biologist, as a test of interest may give a statistically significant result only when a smaller number of tests are performed. If a result is no longer ‘significant’ after adjustment, that is not to say that it is not a true biological result. I have tried to mitigate the impact of multiple testing adjustment by only investigating the genes of interest (i.e. the *pir*s) and using the suggested/inbuilt filters of DESeq2 software to independently remove *pir*s which have low baseline values (which will likely be of no consequence to the analysis beyond creating extra tests for adjustment) (Kozak & Powers, 2017; Love et al., 2014).

I removed one *pir* (*PBANKA_0524600*) from analysis that had been annotated in the genome as a member of the gene family (as of PlasmoDB v48), however I was not convinced that it was sufficiently related to the other family members to be considered such. Not only had this *pir* previously escaped sub-family classification due to its divergent nature, it also only had two exons instead of the normal three. It is probably best understood as *pir*-like since it is certainly closely related to the family, as confirmed by the presence of part of the crystal structure of the PIR domain by Phyre2 (Harrison et al., 2020; Kelley et al., 2015).

One of the starkest results from initial overviews of the life-cycle data was the domination of the transcriptome by a single *pir*, the ancestral sequence. Indeed, close examination of RNAseq studies of *P. c. chabaudi* AS (Brugat et al., 2017; Spence et al., 2013) and single-cell RNAseq of *P. berghei* (Howick et al., 2019), demonstrate the same gene/ortholog as being among the most highly transcribed *pir*s across samples, although it is not discussed by these publications. I explored the hypothesis that this gene and its orthologs are present and highly transcribed across disparate *Plasmodium spp.* which contain *pir*s. Every recent genome with an annotated *pir* family appears to have an ortholog of this gene, with only the most divergent of species absent for it (avian malarias, which only have some *pir-like* genes (Böhme et al., 2018), and the Laverania sub-genus which only have the similar *rifin/stevors* multigenes which may not be related to the *pir*s (Harrison et al., 2020)). This continues to suggest that this sequence is the one from which the other members of the family are derived, expanding through gene conversion/duplication and developing divergent sequence over time (Frech & Chen, 2013). The gene must retain an important function given that it is so highly conserved and slow to mutate - as shown by the 100% conservation between *P. c. chabaudi* AS and CB strains. It is interesting to ponder whether the last common ancestor of these species possessed only one copy of the ancient *pir* gene, or whether the gene had started duplicating into a gene family before the lineages diverged. Why would evolution favour the spawning of such a large family of closely-related genes? Do they provide redundancy, novel functionalities, both, or neither? The high transcription across most of the *Plasmodium spp.* datasets (those from gametocytes and asexual blood stages at least) suggest that this *pir* gene is important across the species that possess it. I believe that this would make a pertinent target for gene localisation and knock-out/down studies in order to learn more about this protein and its function. Although its distinctive expression levels and sequence may mean it may have a different function from the rest of the family, the fact that it likely gave rise to the family suggests that the function of the ancestor may at least be similar to its descendants.

Chapter 4. Results - Transcription of the *pir* genes across the intraerythrocytic development cycle of *P. c. chabaudi* AS

4.1 Introduction

Although the asexual cycle of *Plasmodium* in the mammalian host is often considered as a single entity when extracting RNA from blood samples, the cycle is highly dynamic with different developmental stages. Each stage has a different role in the continuation of infection and the manifestation of symptoms; merozoites invade red blood cells, “ring” stages parasites begin hijacking host cellular machinery, trophozoites initiate proliferation, and schizonts can bind to endothelial cells and can cause clinical symptoms through haemolysis and localised capillary blockage/inflammation. Transcriptome analyses at time points of development across the intraerythrocytic development cycle (IDC), have found major changes of many genes over its progression. *P. falciparum* microarray analysis demonstrated that around 60% of all identified genes are transcribed during the asexual cycle, showing the extensive transcriptional activity of this stage of development (Bozdech et al., 2003). Three-quarters of these transcribed genes has a single major peak of transcription (based on the authors’ definitions and Fourier transformations), showing an mRNA periodicity substantially stronger than that of fission yeast (Spellman et al., 1998), which emphasises the precise timing of transcriptional regulation. Many of these co-transcribed genes have similar functions to one another, suggesting tight epigenetic control over cellular operations. Similarly, *P. vivax* microarray analysis revealed that over 3500 genes have peaks of transcription at given time points which can be double their lowest phases of expression (Bozdech et al., 2008). These transcriptional profiles can be ordered into neat ‘phaseograms’ of normalised peak/minimal transcription values for individual genes spread across the asexual cycle (Bozdech et al., 2003).

In this context, it is clear that samples from single time-points during the asexual cycle, or from ‘mixed’ combinations of the stages, do not fully capture the diverse transcriptional profiles of the parasites in the blood. Often single time-point measurements are necessary because of the asynchrony of malaria erythrocytic cycles in vivo in the blood or *in vitro* cultures, when *Plasmodium* quickly loses any natural

synchronicity. Although synchronicity can be enforced through treatments, such as the chemical lysis of mature cells (Lambros & Vanderberg, 1979), this can add another laborious step in the production of parasite RNA, with potentially artefactual or population-depleting consequences. The rodent malaria parasites, *P. berghei* and *P. yoelii* exhibit asynchronous development even during *in vivo* erythrocytic cycle in rodents (Cambie et al., 1990; Thurston, 1951). Indeed, the exact periodicity of the *P. berghei* asexual cycle in mice has vexed researchers since it was first investigated, as microscopy of blood films at any given time showed all blood-borne stages of development (Thurston, 1951; van den Berghe et al., 1950). Since *P. berghei* has asynchronous infections in laboratory mice, the existing RNAseq data does not cover individual stages of development for the asexual cycle, except for synchronised *in vitro* culture (Janse & Waters, 1995; Modrzynska et al., 2017; Otto et al., 2014) and single-cell analysis (Howick et al., 2019; Reid et al., 2018). In order to obtain bulk RNAseq of the *in vivo* asexual cycle I turned to the closely-related species *P. chabaudi chabaudi,* AS strain.

P. c. chabaudi AS is another *Plasmodium* species infecting rodents, which was discovered in African thicket rats in the Central African Republic by Irene Landau (Landau, 1965). This species is generally less virulent than *P. berghei* in laboratory mice, causing symptoms such as anaemia, weight loss and hypothermia, but seldom killing the animal (Bafort, 1968). After reduction of the acute blood blood-stage parasitaemia, the parasite recrudesces, leading to the establishment of the chronic infection (Yoeli et al., 1966), in contrast to *P. berghei*, which is too virulent to reach a chronic stage in normal laboratory strains of mice, such as C57Bl/6. Importantly for my study, *P. c. chabaudi* AS has 211 *pir*s in its genome and a synchronous asexual cycle in the circulation of mice (Hawking et al., 1972; Otto et al., 2014). The cycle lasts 24h, and under a natural day-night schedule schizogony and merozoite invasion occur around midnight (Killick-Kendrick & Peters, 1978). The *P. c. chabaudi* model has been useful for the investigation of sequestration and rosetting of mature iRBCs (Cox et al., 1987), the role of inflammation in cerebral malaria independently of brain sequestration (Sanni et al., 2004), the interactions between the parasite and host circadian rhythms (Subudhi et al., 2020), and drug resistance (Cowman & Lew, 1989) (uses in research reviewed by (R. Stephens et al., 2012)). In particular, *P. c. chabaudi* has been employed to investigate the immune response to malaria parasites, such as the role of regulatory cytokines like IL-10 in mediating pathology (Linke et al., 1996), the importance of B-cells

for clearance of infection (von der Weid et al., 1996), the characterisation of atypical memory B cells (Pérez-Mazliah et al., 2018), and the immune controls of Th1 responses to acute and Th2 responses to chronic infection (Langhorne et al., 1989, 1998; Linke et al., 1996).

This species has already been used in studies of the role of the *pir* genes (as discussed in more detail in Chapter 1). In summary, by comparing the virulent SBP and the less-virulent MT methods of infection, it was found that the mosquito vector did not just impact the severity of the disease but also the diversity of *pir*s transcribed, as the transcriptome of SBP *P. c. chabaudi* was dominated by a singular *pir* mRNA (the SBP-high *pir*) and that of MT parasites more broadly spread out across the repertoire (Spence et al., 2013). The related strain *P. c. chabaudi* CB also exhibits this difference, in a model that is generally more virulent to laboratory mice (Lin et al., 2018). A similar pattern was seen with acute and chronic infections, the latter of which caused more pathogenic infections (when inoculated into nonimmune mice), and similarly to SBP parasites, strongly transcribed only a few *pir*s compared with parasites of the acute population (Brugat et al., 2017). These hint towards a contribution of *pir*s to virulence and how this may link with parasite hedge-betting are some of the most tantalising links to *pir* function that we have. Like many studies, the above investigations used a single time point in the erythrocytic cycle; in this case the trophozoite stage, providing a limited view of *pir* transcription in *P. c. chabaudi*. How the different conditions impact *pir* transcriptional patterns across the development of the asexual blood stages remains an open question. It is unknown whether the diverse *pir* transcription in MT parasites or the restricted *pir* transcription of SBP parasites is observed across the entire 24h cycle or just at the single time measured. Furthermore, the timing of the global and *pir* transcriptome of the parasites may be affected by serial blood passage of *P. c. chabaudi*, thus rendering SBP parasites a poor reflection of overall *pir* transcription in the more natural blood-stage infection. The studies *of pir* genes in *P. c. chabaudi* (Brugat et al., 2017; Lawton et al., 2012; Spence et al., 2013), as well as studies from other species (e.g. Otto et al., 2014, and Zhu et al., 2016) show that *pir*s are transcribed highly overall in the asexual blood stages, even in the SBP/chronic infections which downregulated *pir*s relative to acute MT. Hence it is reasonable to suggest that the PIR proteins may play an important role in the asexual blood cycle. However, an open question remains as to whether they are required throughout the entire cycle or just at particular times, and which different sub-families of *pir* genes need to be transcribed at different points of development.

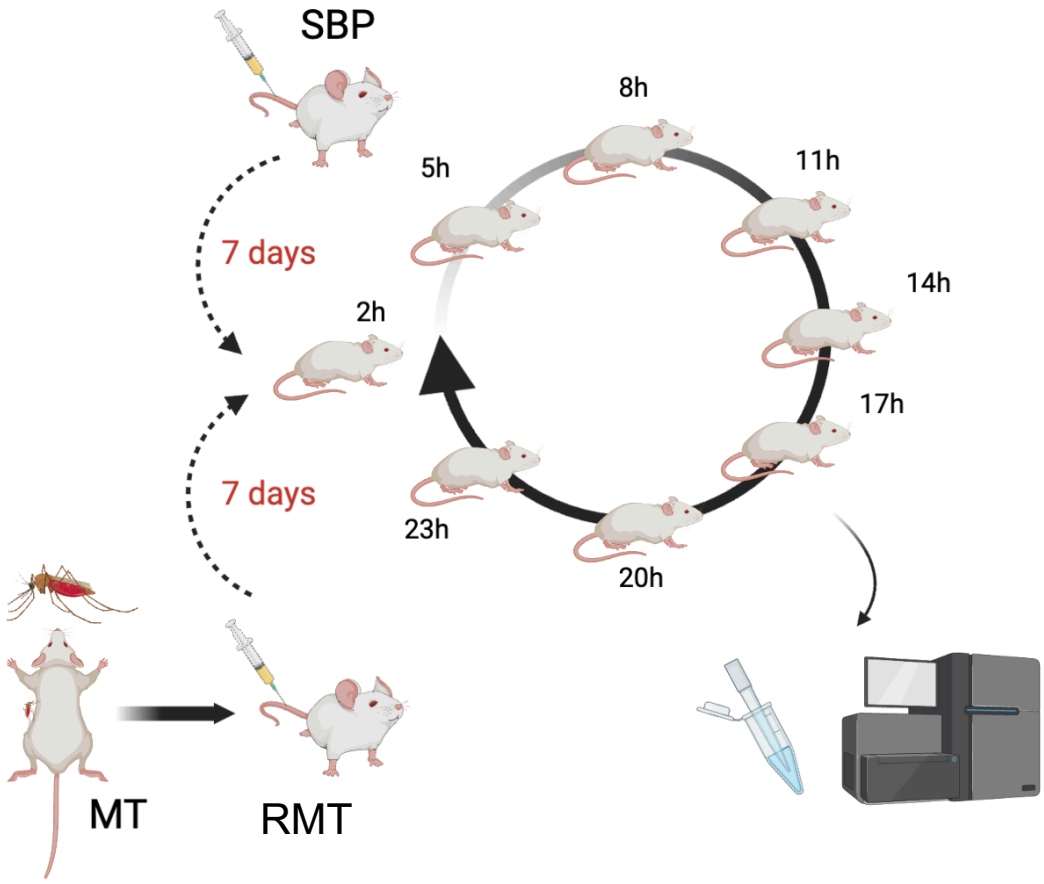


Figure 4.1 – Schematic of the RNAseq extraction across the asexual life cycle for both SBP and RMT infections of *P. c. chabaudi* AS.
Figure drawn using BioRender (https://app.biorender.com).

Microarray data of RNA levels of 49 *pir* genes across the IDC showed that they are transcribed at different times across the cycle (Lawton et al., 2012). Immunofluorescence has indeed demonstrated the expression of some PIR proteins across many *P. c chabaudi* trophozoites and merozoites, however not in all the cells observed (Giorgalli et al., 2022; Yam et al., 2016). Zhu and colleagues (Zhu et al., 2016) showed in *P. vivax* that the *pir* transcriptional profile varies across the asexual cycle, peaking around merozoite development. Data from *P. berghei,* although limited to three time points corresponding to the rings, trophozoites and schizonts, also showed a change of transcription with time – peaking in rings but lowest in schizonts (Otto et al., 2014). Both these datasets are from parasites grown *in vitro*, which can vary substantially from those developing *in vivo*, as in the case of *P. falciparum* (reviewed by (LeRoux et al., 2009)). These differences can clearly be seen in the transcriptional

patterns of multigene families such as *var,* and the *SICAvar* in *P. knowlesi* (Bachmann et al., 2016; Lapp et al., 2015). However, the *P. berghei* and *P. vivax* samples had been minimally cultured (could be described as *ex vivo* culture – fresh sample cultured for one cycle or less outside the host) as these species have no long-term *in vitro* system, meaning that they would have had little time to adapt compared to biologically relevant conditions. Notwithstanding this, an ideal model would be entirely *in vivo*.

Hence *P. c. chabaudi* provides the opportunity to investigate synchronised stages of asexual development, *in vivo*, to compare the *pir* transcription across the erythrocytic cycle, and investigate the interplay of transcription with different severities of infection. Experiments conducted by others in the Langhorne lab (see Materials and Methods) have extracted RNAseq of asexual blood stages across a twenty-four-hour period every three hours, from a recently mosquito-transmitted (RMT) infection (one passage since MT – see Materials and Methods). A similar time course was performed with an SBP infection included for comparison as a more virulent and synchronous model, albeit one less relevant for real-world conditions (see Figure 4.1). In addition, the RMT cells were cultured *ex vivo* briefly to achieve schizont samples, as these stages are absent from the circulatory samples taken from these infections. Using these data, I have investigated the global and *pir* transcriptome across the *P. c. chabaudi* AS asexual cycle, and how this varies based on transmission method and the *pir* sub-family.

4.2 Aims

- Evaluate *pir* gene transcription changes across the asexual blood cycle of *P. c. chabaudi* AS.
- Does transcription over time depend on the *pir* subfamily or genomic locus?
- Does serial blood passage (SBP) of *P .c. chabaudi* AS affect the transcriptional profile of *pirs* over the 24h cycle?
- Do the *P. c. chabaudi pir* transcriptional changes during the asexual blood cycle match those seen with the limited *P. berghei* data available?
- Predict when given members or groups of the *pir*s are needed during the asexual development cycle.

4.3 Results

4.3.1 RMT samples are mildly asynchronous but still dominated by expected stages of the IDC

By counting the stages corresponding to the RNAseq time-points via light microscopy (Figure 4.2A) the dynamics of RMT asexual development can be observed; with ring forms becoming trophozoites (midpoint at 5h), trophozoite numbers peaking at 11h, followed by trophozoites becoming depleted and rings dominating around 17h post-invasion. Schizonts remain negligible at all time-points as they bind the endothelium in the tissues and organs and will not be seen in the circulation. These quantifications are supported by the RNAseq-based deconvolution (using 1-to-1 orthologues in scRNASeq data of *P. berghei*) in Figure 4.2B. The major disparities between the light microscopy counting and deconvolution are the presence of merozoites and schizonts at 14/17h, and female gametocytes at 17/20h. The merozoite differences can be explained by the fact that merozoites cannot be quantified by light microscopy, and the detection of schizonts and female gametocytes in the deconvolution could be due to marginal cases that were too immature to be defined by microscopy but had sufficient change of gene expression to be classified by the deconvolution tool. Based on Figure 4.2 there were no time-points of the RMT infection where there were ‘pure’ extractions of one stage of development. This is to be expected as *P. c. chabaudi* AS infections initiated by mosquito inoculation are slightly asynchronous. Hence, the time-points of stage transitions can blur into one another. As expected, the SBP stage composition is more synchronous than that of RMT, but also has slightly shifted times of stage transition, as discussed in more detail at the end of this section.

I investigated global differences between the transcriptomes of the samples, including the RMT schizont cultures (sch.culture), using Principal Component analysis (PCA). The first PCA plot (Figure 4.3A) demonstrated that the RMT and SBP samples were distinct from one another along PC2. The RMT samples also did not appear to form an expected cyclical structure. A major reason for this was a single pseudogene *PCHAS_0420800*, located at the very end of the assembly of chromosome 4. Figure 4.3B shows the TPM values of this pseudogene across the time-points.

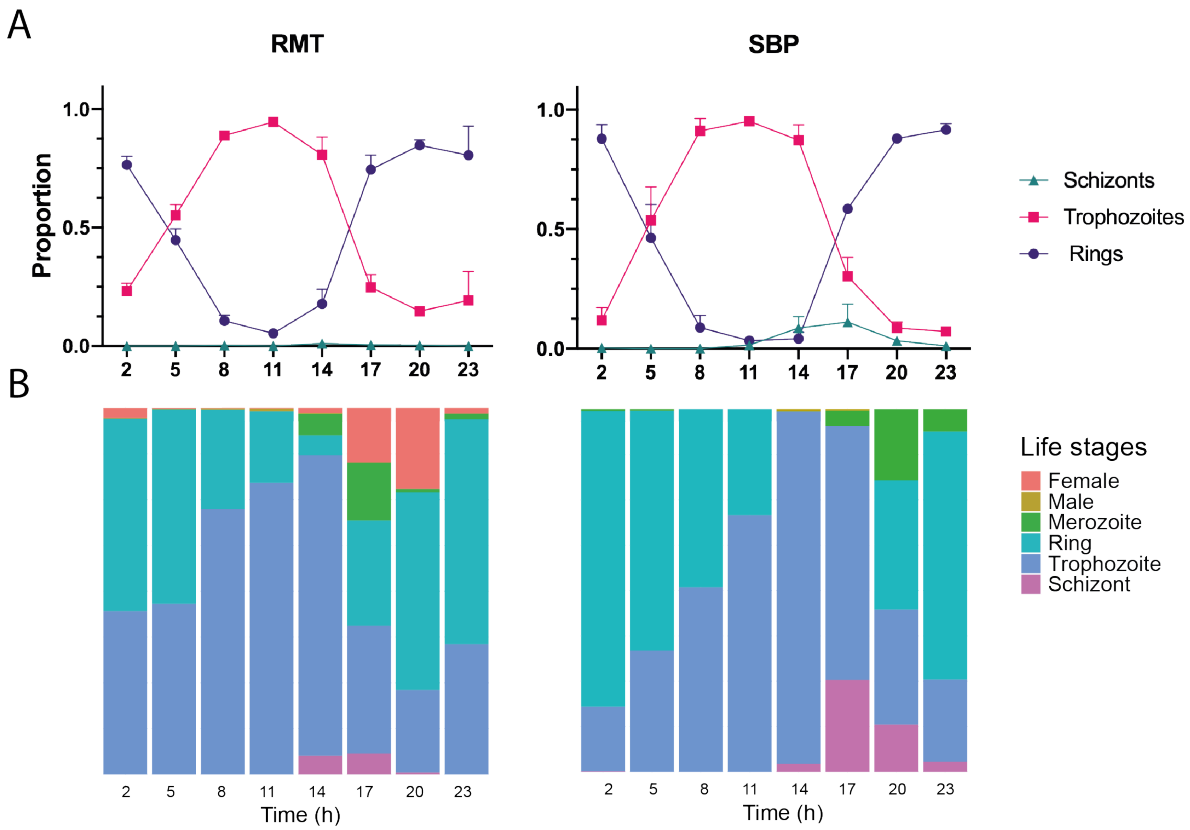


Figure 4.2 – Proportions of parasite blood stages for each timepoint.
A. The differential counts of each stage of the asexual blood cycle (Rings, Trophozoites and Schizonts) for each transmission method, enumerated through light microscopy by Deirdre Cunningham. They are shown as proportions of total iRBC. B. Deconvolution of the RNAseq data to identify what proportion of the reads are explained by different stages of the life cycle. This used 1-to-1 orthologues between *P. c. chabaudi* AS and *P. berghei* single cell data (Howick et al., 2019)**. Female and Male refers to the respective sex of gametocyte. Conducted by Adam Reid.**

There is a disparity of transcription between RMT and SBP samples, and in some RMT RNAseq data this singular RNA accounts for over 0.6% of the total TPM. With such a large contribution by one, possibly defective, gene, I decided to investigate the variability between the samples when this gene was removed and the TPM data re-normalized. As Figure 4.3C demonstrates this makes a pivotal difference to the second biggest component of variation in the samples as they now do not separate by type of infection. The transition across time points during the RMT asexual cycle is much clearer, although the SBP 20h and RMT schizont cultures remain outliers. For the rest of this analysis, I have used the re-normalized TPM data without *PCHAS_0420800.*

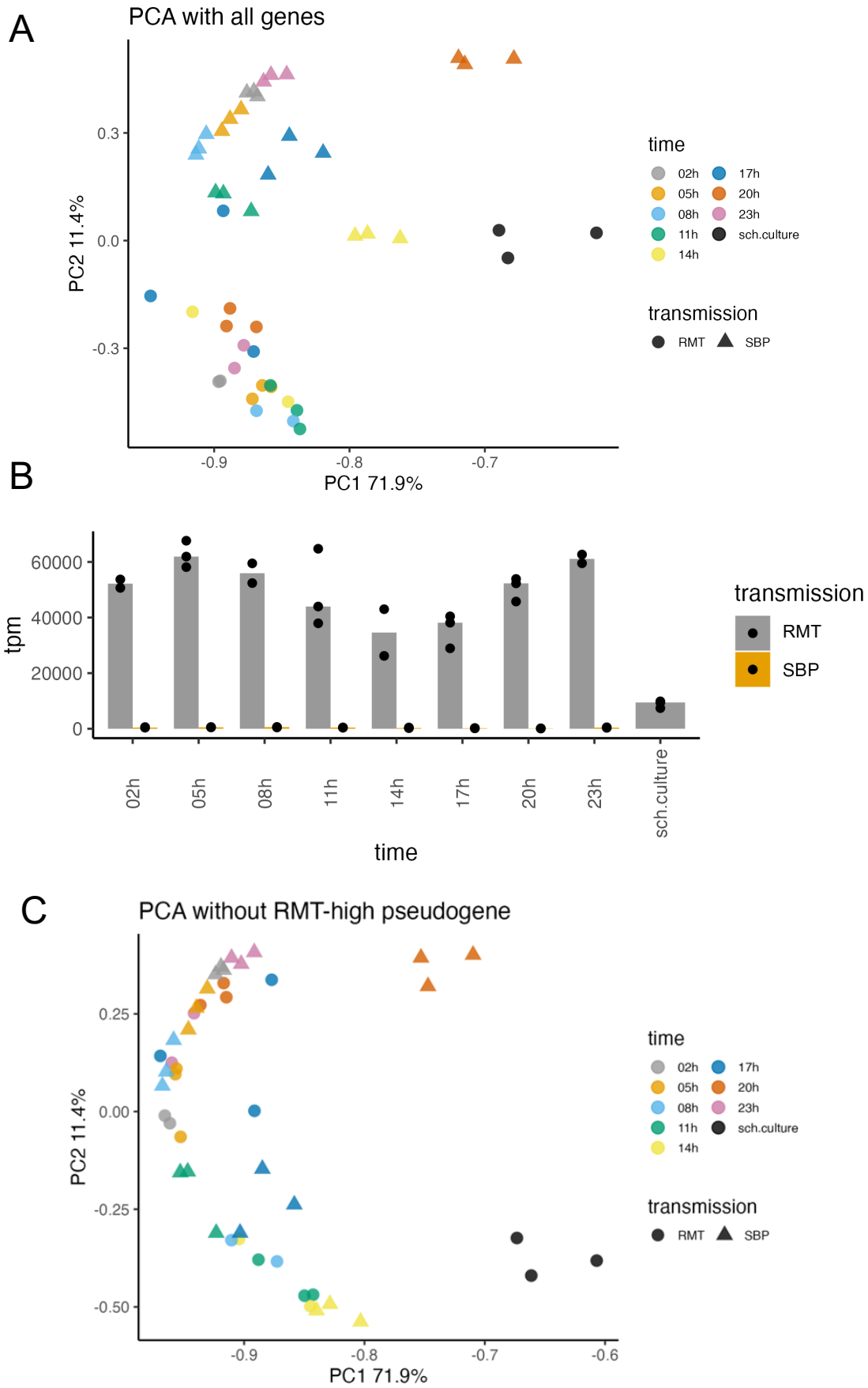


Figure 4.3 – Variation within and between the RMT/SBP asexual blood cycle samples.

A. PCA plot of the RNAseq samples using all gene TPM values. Shape is based on transmission method and coloured by the time-point of extraction/schizont culture. B. Bar chart of the TPM values for the pseudogene *PCHAS_0420800*, which is transcribed much higher in RMT samples than SBP. C. PCA plot of the RNAseq samples using the gene TPM values without the RMT-high pseudogene (re-normalised to sum to 1e6). Both PCAs were scaled to unit variance and zero-centred.

By calculating z-scores across the twenty-four-hour time points I have been able to observe the variation of global gene transcription (see Figure 4.4A). By ordering the heatmaps based on when genes are transcribed at their maximum a highly ordered pattern of gene transcription can be observed, as has been with other *Plasmodium sp.* IDCs (Bozdech et al., 2003, 2008). There are major times of upregulation at 5h, 14h, and 17h of the life cycle (corresponding generally with periods of cell stage transition: rings-to-trophozoites, trophozoites-to-schizonts, and possibly merozoite-to-early rings).

Although both RMT and SBP have similar asexual blood cycle rhythms to one another, there are differences which must be addressed before they can be compared. For example, the return of ring stages in the blood films begins earlier in RMT samples (14h), and the SBP samples have a higher proportion of schizonts, peaking near 10% at 17h post-invasion (Figure 4.2A). This is also reflected in the deconvolution results (Figure 4.2B), which show that the appearance of merozoites occurs earlier (14-17h) in RMT samples than SBP (17-20h). From Figure 4.4A and B it can be observed that there are swathes of genes which peak in transcription at the same time in infections initiated by both methods of transmission. However, there are also many genes which are ‘out of sync’ when comparing RMT and SBP, usually peaking earlier/later by one time-point. It is unclear whether this is a biological effect or an artefact of experimentation (such as variation between parasite clones/mice/researchers). These shifts may be attributable to the differences in the proportions of different asexual stages found across the twenty-four-hour cycle. Given that the underlying causes cannot be evaluated by this study I noted that direct comparison between time-points of RMT and SBP may be impacted by the altered proportions of IDC stages in each infection.

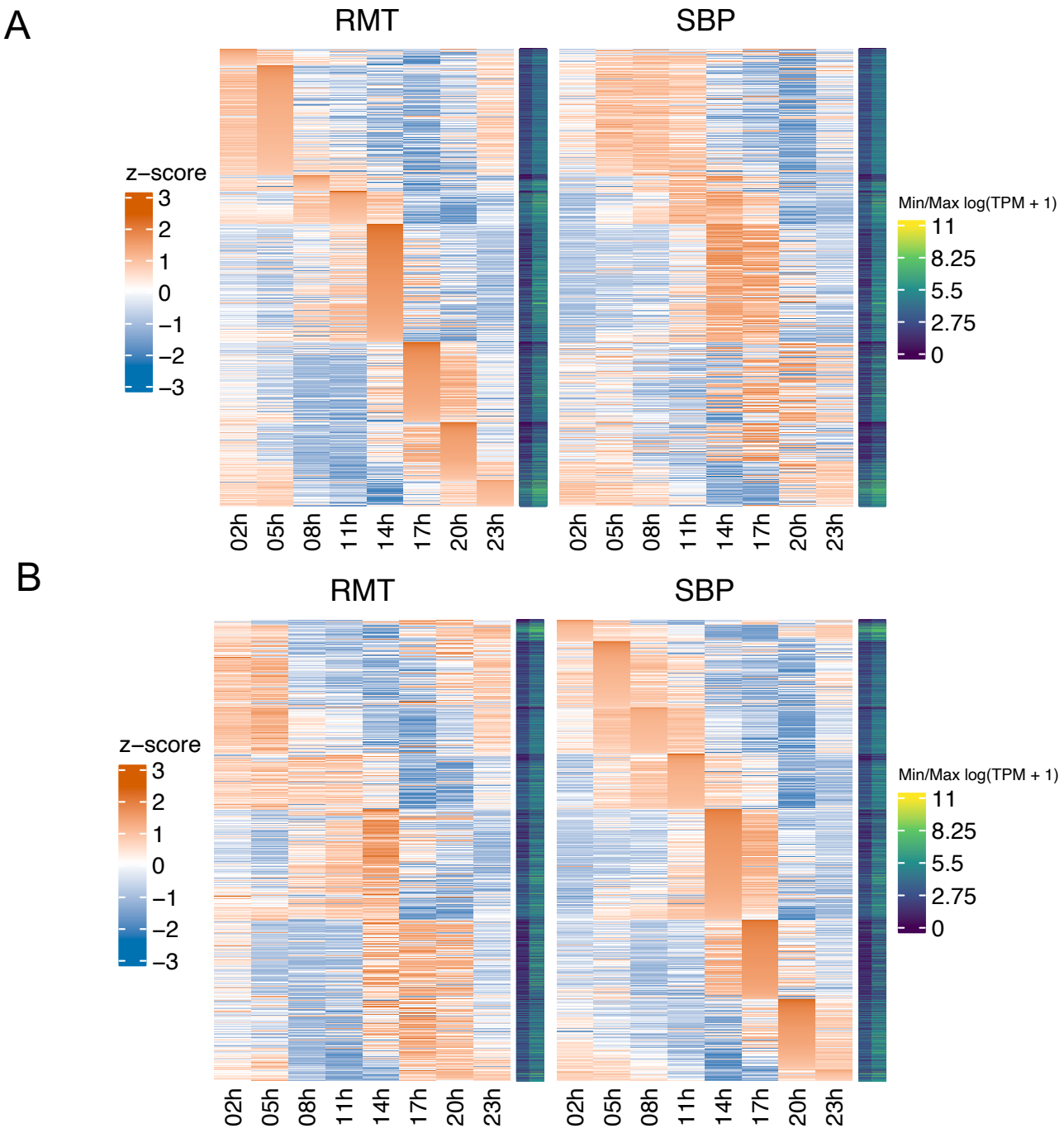


Figure 4.4. Phaseograms of maximum/minimum transcription of every gene of the transcriptome across the twenty-four-hour asexual life cycle of *P. c. chabaudi*. Main heatmaps show the z-scores of genes; a measure of the variation of the transcription of a gene, with the stages of maximal transcription having higher z-scores and minimum more negative z-scores. More consistent transcription across time points will lead to lower absolute values of the z-scores. Z-scores were calculated for the RMT (without the schizont culture samples) and SBP experiments

independently using the averaged TPM values. The heatmaps are ordered by which genes are transcribed at their maximum level (most positive z-scores) from 02h to 23h in RMT (A) and SBP (B). Side heatmaps show the minimum/maximum log(TPM + 1) from across the time-points for comparison with the absolute expression levels. Note that high value z-scores do not reflect high transcription, just high variation of transcription.

4.3.2 Transcription of *pir* genes peaks during the later trophozoite stages

I next sought to investigate how *pir* transcription changes during asexual blood development. Using PCA again, except this time just with the *pir* genes, I created an overview of the variation of the *pir* profile with time (see Figure 4.5). This shows that the *pir* transcriptome is diverse across the IDC, with a progression of transcription as the parasite cells develop and then complete the cycle. This is apparent across both components of variation, covering 91.5% of total variation of *pir* gene transcription. However, the SBP samples are clearly separated from the RMT time-points, mostly by PC2 (11.7% of the variation). Additionally, the majority of variation is among the RMT samples, which are far from one another while the SBP samples form a tight cluster, which shows that the RMT parasite populations have more heterogeneous *pir* profiles.

By analysing total *pir* gene transcription levels across the asexual life cycle, a consistent peak at the 14h time-point can be observed, during the late trophozoite stage of development in *P. c. chabaudi* (see Figure 4.6). This is the same between RMT and SBP transmission despite the reduced transcription of *pir* genes overall in SBP. After the RMT peak, there is an abrupt drop in *pir* transcript abundance (at 17h). This may be explained by this moment being a time of trophozoite-to-schizont transition, and the RMT samples not capturing the sequestering schizont stage effectively (Figure 4.2). The *in vitro* cultured schizont samples suggest that these stages are indeed enriched for *pir* gene transcription compared to the parasites in the 17h circulation. This is an example of how the cultured schizonts can help to plug the ‘schizont gap’ in the RMT time course. Further evidence for this explanation comes from the SBP transcriptional *pir* profile, which has non-sequestered schizonts in the samples taken and shows no such major drop after the 14h peak. Overall, there is a cyclical pattern of transcription, with lower transcription when ring stages predominate at the first and last time-points, and a peak during late trophozoites.

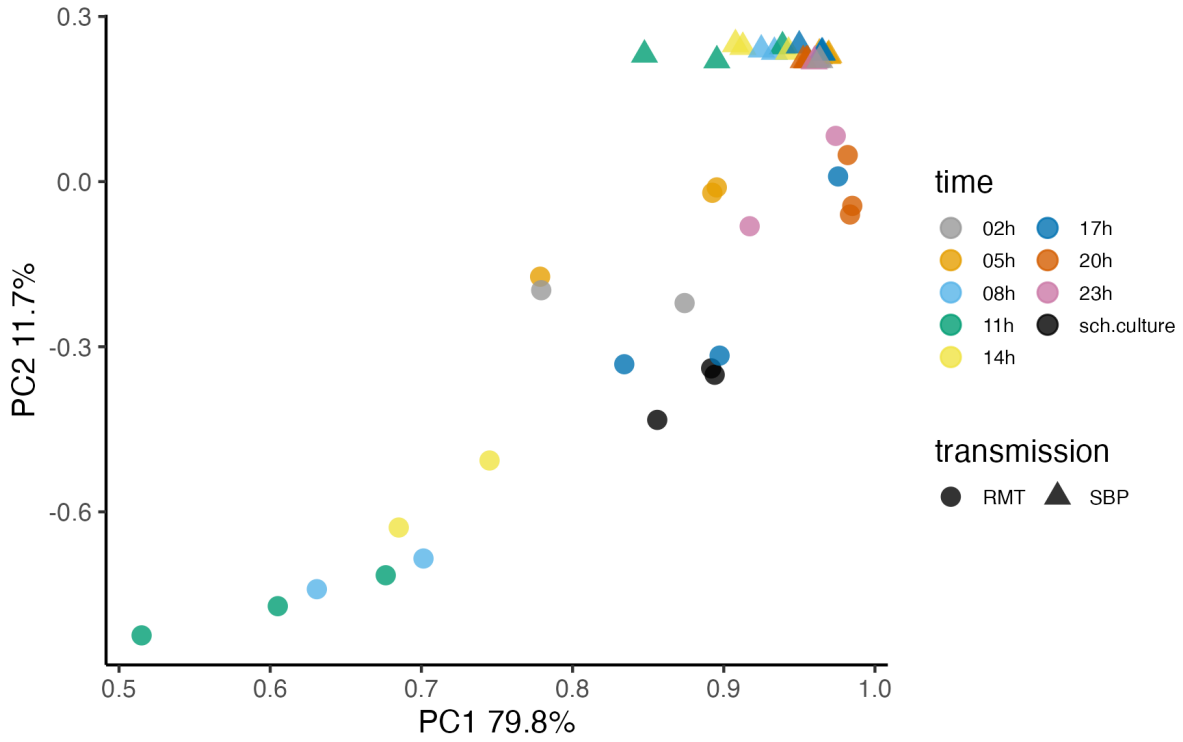


Figure 4.5. PCA plot of the RNAseq samples using the *pir* gene TPM values. Shape based on transmission method and coloured by the time-point of extraction/schizont culture. PCAs were scaled to unit variance and zero-centred, so the TPM data did not need to be re-normalised beforehand.

As observed with *P. berghei* and other *Plasmodium spp.*, the ancestral *pir* (*PCHAS_0101200*) is transcribed during the asexual cycle and is often the most highly transcribed individual *pir*. Plotting this gene on its own (see Figure 4.7) I can clearly see that transcription peaks with most of the other *pir*s in RMT, at 14h. Once again, this peak may be influenced by the 'schizont gap', as RMT schizont transcription of the ancestral *pir* is high and the *pir* is transcribed highest at 17h in the SBP parasites. It is possible that RMT would show the same peak at 17h were it plausible to extract the sequestered schizonts alongside the circulating parasites for RNAseq. Throughout the asexual cycle transcription of this gene remains high (relative to other *pir*s) reaffirming the potential importance of this conserved gene in the parasite. I also looked directly at the transcription of the SBP-high *pir* gene *PCHAS_1100300* (see Figure 4.8), as Spence

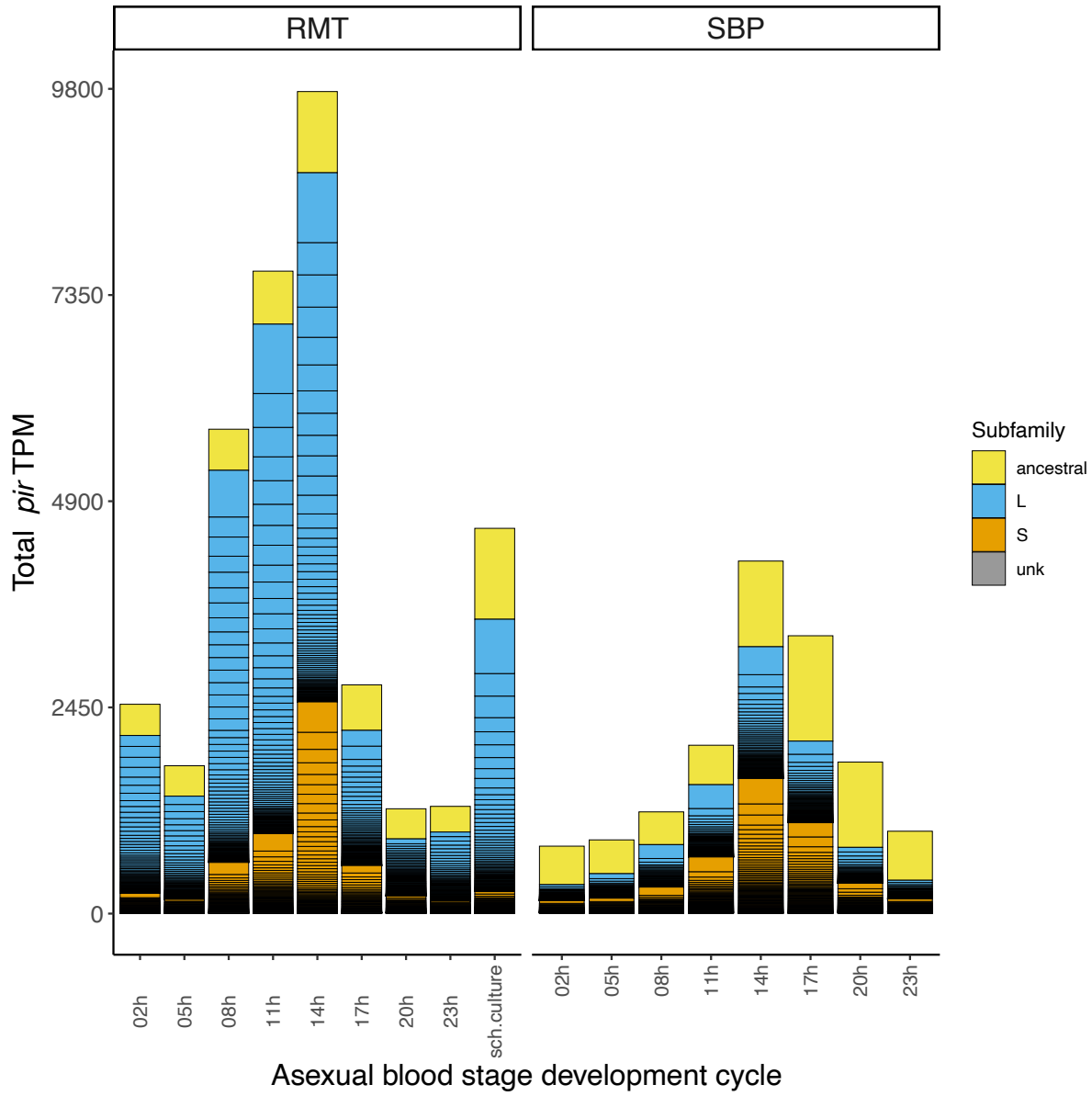


Figure 4.6 – Bar chart of *pir* gene TPM across the asexual cycle and RMT schizont cultures of *P. c. chabaudi* AS transmitted by RMT or SBP.
Each 'stack' of the bar represents the TPM of a single gene. The stacks are sorted first by sub-family (the colours) and then by TPM value, so stacks in the same position on different bars do not necessarily represent the same gene. Genes were filtered by only including those transcribed above 1 TPM in at least one sample.

et al, 2013 found that this *pir* was transcribed at much higher levels in SBP samples compared to MT samples. In contrast to the results obtained previously, in these infections the transcription of this gene looks similar between both RMT and SBP at most time-points. Perhaps this is because a population of parasites transcribing this *pir* were selected for by the conditions of the Spence et al, 2013, experiment, but those

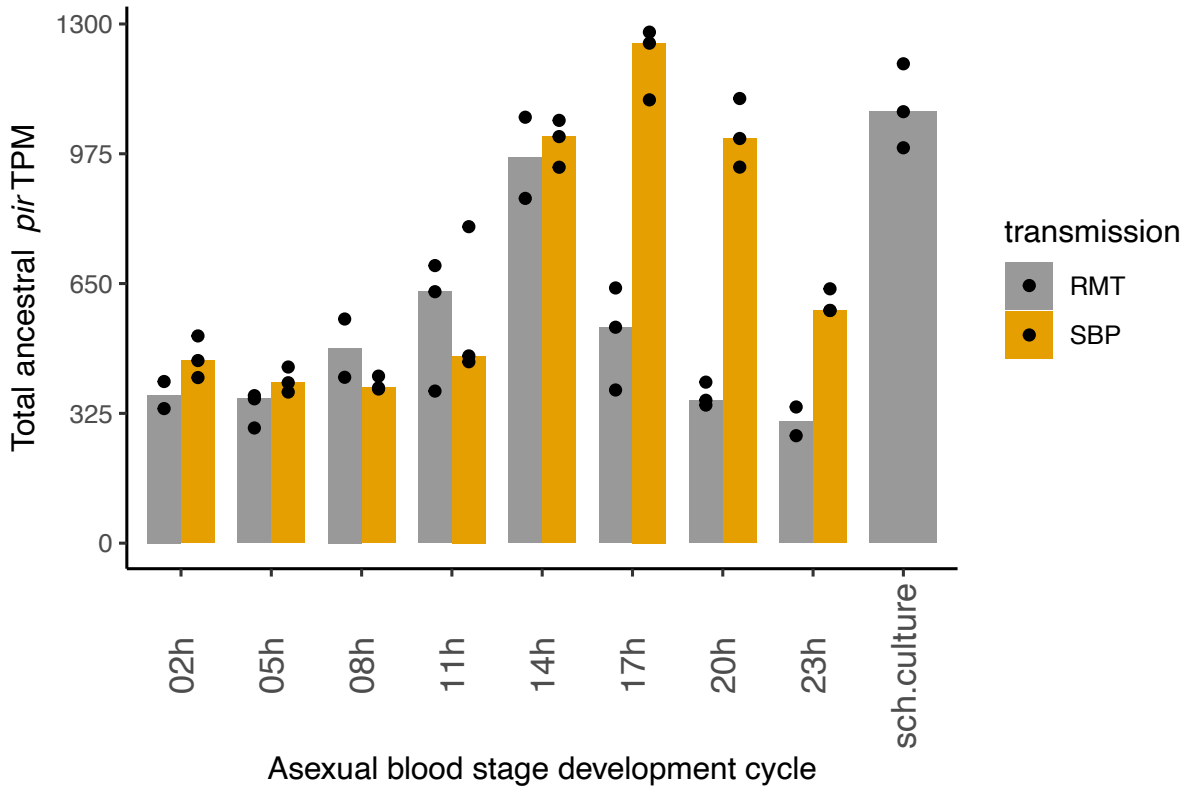


Figure 4.7 – Bar chart of TPM of the ancestral *pir* gene *PCHAS_0101200* across the asexual blood cycle and RMT schizont cultures.
Each dot represents a biological replicate (n=2-3) and colours show the method of transmission. Bars show median TPM.

conditions are not in place for these later experiments. Alternatively, the SBP-high *pir* may now be part of a stochastic broad repertoire of *pir*s transcribed in the RMT infection. These results demonstrate the same patterns that Spence et al., 2013 observed, although the individual genes involved may change even within these laboratory isolates.

To observe the peak of expression for every *pir*, I made similar phaseograms to those above to compare when maximum transcription occurs for the *pir*s (see Figure 4.9). While most *pir* genes indeed peak at the 14h time-point, there are many which peak at other time points. The z-scores demonstrate that multiple *pir*s peak in transcription as early as at the 8h time-point, especially some of those from the Long-form clade. It is important to consider that transcription for a gene may still be high before and/or after the peak itself, as comparison between Figure 4.6 and Figure 4.9 demonstrates that the

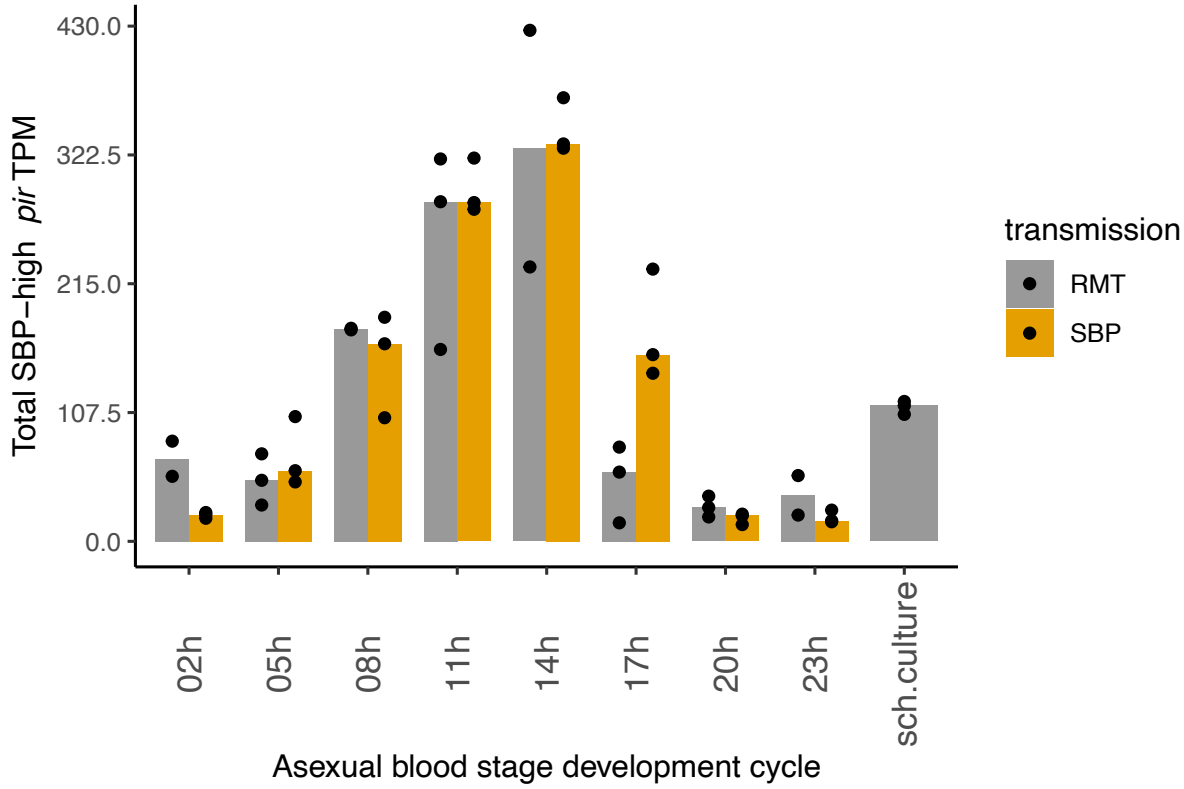
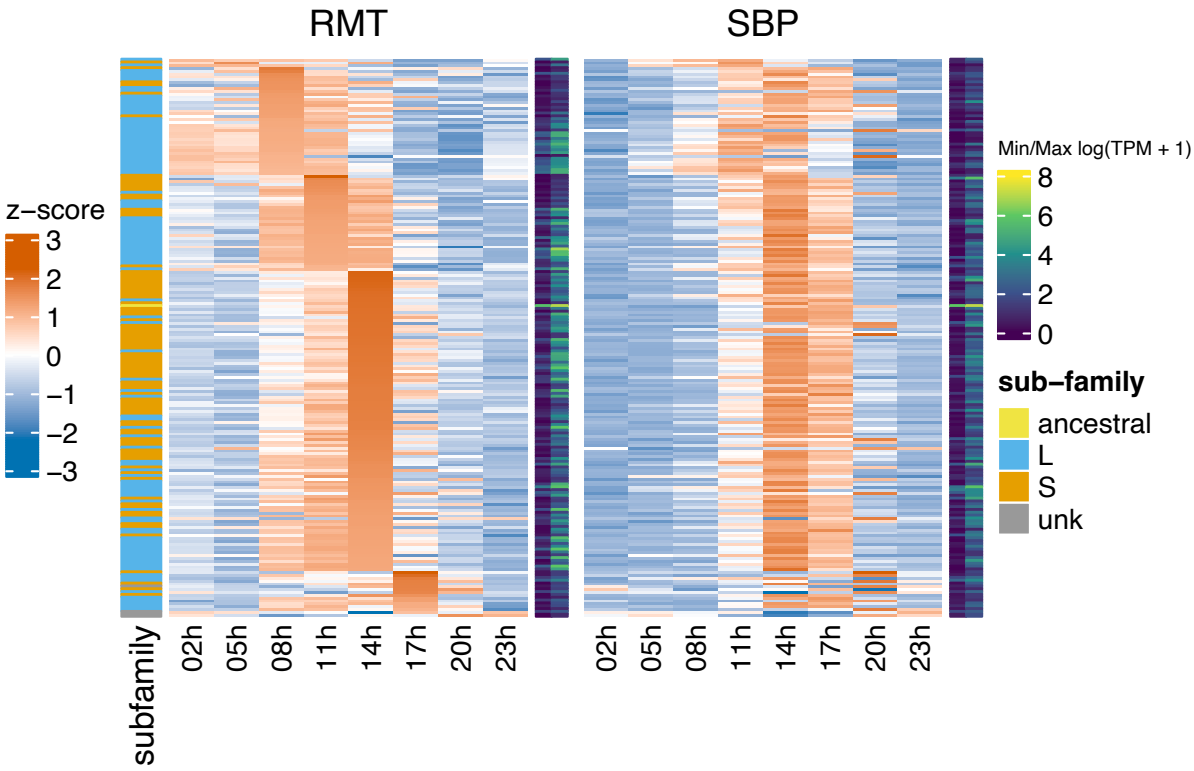


Figure 4.8 – Bar chart of TPM of the SBP-high *pir* gene *PCHAS_1100300* (an individual *pir* massively upregulated in SBP samples compared to MT in Spence et al, 2013) across the asexual blood cycle and RMT schizont cultures.
Each dot represents a biological replicate (n=2-3) and colours show the method of transmission. Bars show median TPM.

L-form *pir*s peak early but are still transcribed during the overall *pir* peak at 14h. Compared to natural infection, blood passaging appears to refine the peak of transcription, perhaps due to increasing the synchrony of the parasites. While most *pir*s peak at 14h, there are multiple which are transcribed highest at other times of the IDC, and in RMT these earlier transcribing *pir*s are mostly L-forms.

4.3.3 Differential transcription patterns of the pir sub-families and the relationship to chronic and acute-associated loci

My investigations above suggest that there is a relationship between the timing of transcription of *pir* genes and the sub-family a *pir* belongs to. To further delineate the relationship between sub-family and transcription I constructed a heatmap of


RMT
SBP
z−score
3
2
1
0
−1
−2
−3
Min/Max log(TPM + 1)
8
6
4
2
0
sub−family
ancestral
L
S
unk
subfamily
02h
05h
08h
11h
14h
17h
20h
23h

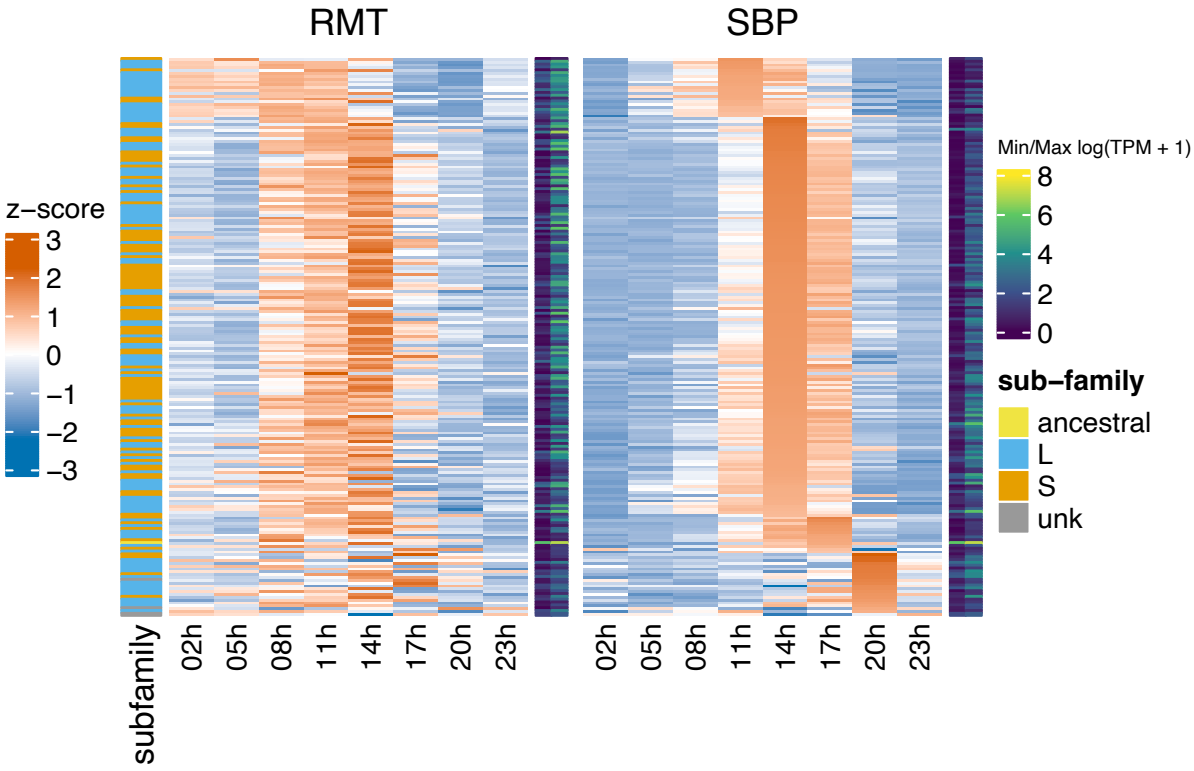

RMT
SBP
z−score
3
2
1
0
−1
−2
−3
Min/Max log(TPM + 1)
8
6
4
2
0
sub−family
ancestral
L
S
unk
subfamily
02h
05h
08h
11h
14h
17h
20h
23h

Figure 4.9 – Phaseograms of the z-scores of the *pir* genes compared between RMT and SBP.
Main heatmaps show the z-scores of *pir* genes; a measure of the variation of the transcription of a gene, with the stages of maximal transcription having higher z-scores and minimum more negative z-scores. More consistent transcription across time points will lead to lower absolute values of the z-scores. Z-scores were calculated for the RMT (without the schizont culture samples) and SBP experiments independently using the averaged TPM values. The heatmaps are ordered by which genes are transcribed at their maximum level (most positive z-scores) from 02h to 23h in RMT (A) and SBP (B). Right heatmaps show the minimum/maximum log(TPM + 1) from across the time-points for comparison with the absolute expression levels. Left heatmap shows the sub-family designation of each *pir*. Note that high value z-scores do not reflect high transcription, just high variation of transcription.

the transcription of the *pir*s split by sub-families within the Long- and Short-forms (see Figure 4.10). For example, the L1 *pir*s in RMT infections begin transcribing 'earlier' in the cycle. This can start at the 02h time-point, corresponding to mature ring stages which have not yet developed into trophozoites, and continuing for several time-points. The L4 *pir*s begin transcribing highly across the 08h-14h time-points, especially among the group of 8 L4s with the highest overall transcription in the sub-family. It is possible that the subsequent dip of L4s at 17h is a consequence of schizont sequestration, as schizont cultures also show high L4 transcription. In contrast, the S-form *pir*s, most clearly observable with the S7s, tend to transcribe acutely at the 14h trophozoite peak. From this, we can see that digging further into the sub-family membership uncovers better detail on how they influence the timing of *pir* transcription.

The *P. c. chabaudi* AS *pir*s can also be grouped into putative functional groups based on their genomic location and relative transcription levels in acute or chronic infections (Brugat et al., 2017). These chronic-associated *pir* loci (ChAPLs) and acute-associated *pir* loci (AAPLs) are closely linked with the sub-families, as the former are enriched for

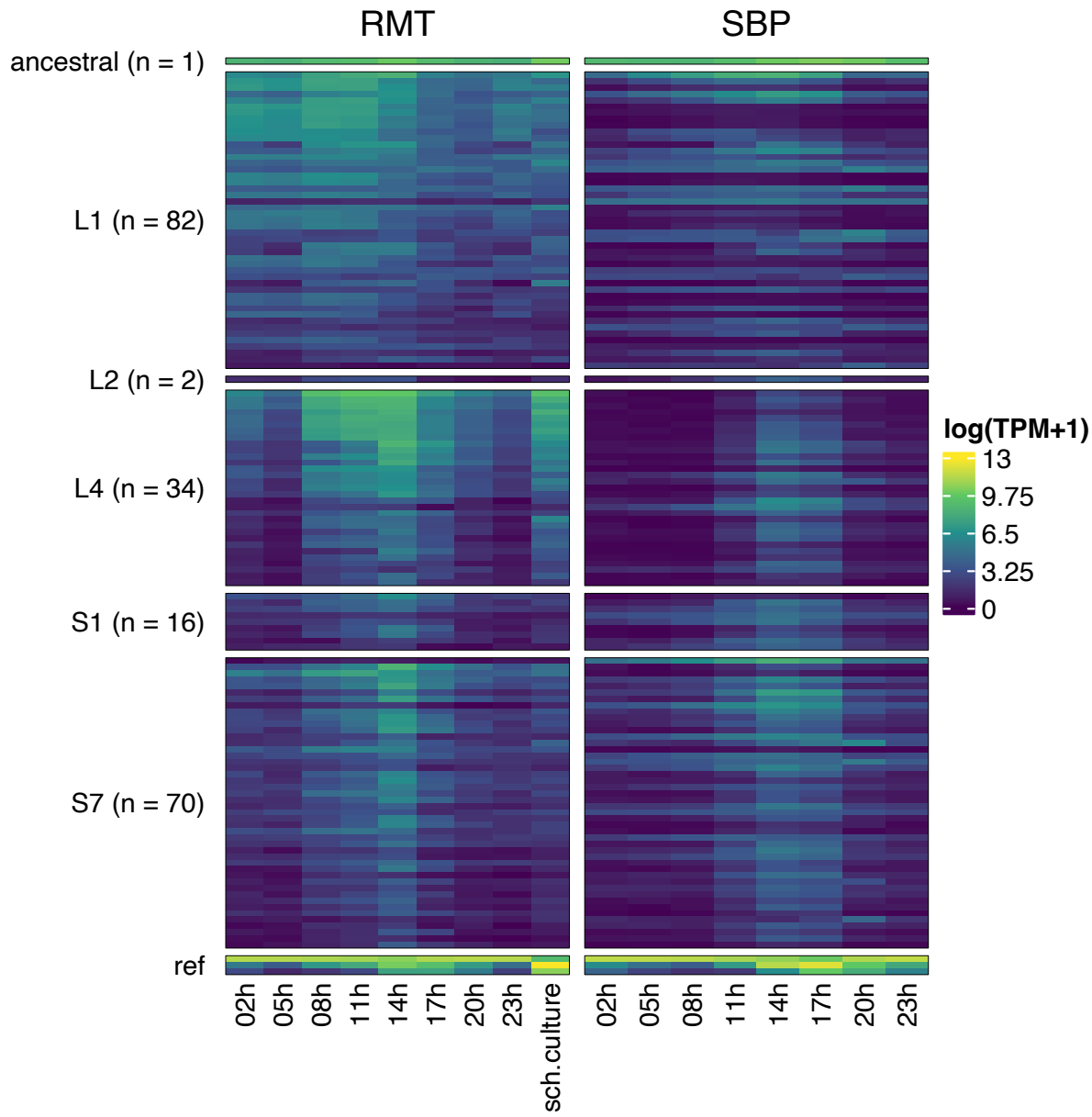


Figure 4.10 – Heatmap of the log(TPM+1) values of the *pir* genes across the twenty-four-hour cycle of *P. c. chabaudi* AS in both RMT and SBP transmitted infections, and RMT schizont cultures.

Genes are grouped and the heatmap split by the sub-family of the *pir* genes. Ref genes are three housekeeping genes for comparison; HSP70, MSP1 and AMA1. Genes were filtered to remove any that were not transcribed above 10 TPM in any time-point average. Sub-family labels on the left also show the number of members of the sub-family, which may not correspond to the number on the heatmap due to the filtering step. Genes in each section of the heatmap are ordered by average transcription across all samples.

S7s and the latter for L1s. Even though this experiment does not include any chronic infections, the transcription of these groups in natural RMT infections is clearly distinguished in the way one would expect from the *pir* sub-families that make them up (see Figure 4.11). The AAPL genes show a distinctive peak of transcription around 14h. The ChAPL genes show a broader timing of transcription across the time-points from as early as 2h to 14h. Sometimes this transcription is so early that the proportion of transcripts is decreasing by 14h, and the rise of transcripts occurs at 23h. By looking at the sub-families it appears that the transcriptional pattern within the ChAPL genes depends on whether a gene is L1 (early transcription from 2-14h) or L4 (transcription from 8-14h). The L4 *pir*s are found in both AAPLs and ChAPLs, however only those in the ChAPLs appear to show such a broad timing of transcription. So, the L4 sub-family itself can be divided into the sharply peaking AAPL-L4s (~14h, like the S-forms) and the broadly peaking ChAPL-L4s (8-14h, like the L1s).

To corroborate the observed patterns, gene set enrichment was used for statistical analysis of the *pir*s and *pir* groupings between consecutive time-points of the asexual cycle (see Figure 4.12). The only significant upregulation of the whole *pir* family in RMT is at 8h after the 5h time-point, while *pir* downregulation occurs at 17h. In contrast to the peak of *pir* transcription at 14h (late trophozoites), the main upregulation of *pir* gene transcription actually occurs at 8h in these infections (maturing from early to mid-trophozoites). The broad transcription of L-form and ChAPL *pir*s, compared to the more acute activation of S-form and AAPL *pir*s, is also evident. The slow decrease of transcription for the L1s after the 14h peak means that there is no statistically significant decrease found from each pairwise comparison across time, suggesting that transcription remains high even if it is changing. The L4s demonstrate a major upregulation at 8h, and then a reduction at 17h that is probably a consequence of the schizont gap (as L4 transcription is high in the schizonts - Figure 4.10). The L4s are split across ChAPLs and AAPLs, and comparing these groups in RMT indicates that the ChAPL L4s are transcribed highly for longer (upregulated at 8h then downregulated at 17h), while the AAPL L4s have a sharper peak (upregulated at 8h, then again at 14h, before sharp downregulation at 17h, similar to the S-forms) (Figure 4.10 and Figure 4.11). One ChAPL locus, on the right-hand arm of chromosome 6, is downregulated early, by the 14h peak of *pir* transcription. This analysis suggests that *pir* upregulation generally occurs at the 8h time-point of the asexual cycle and downregulation by the 17h

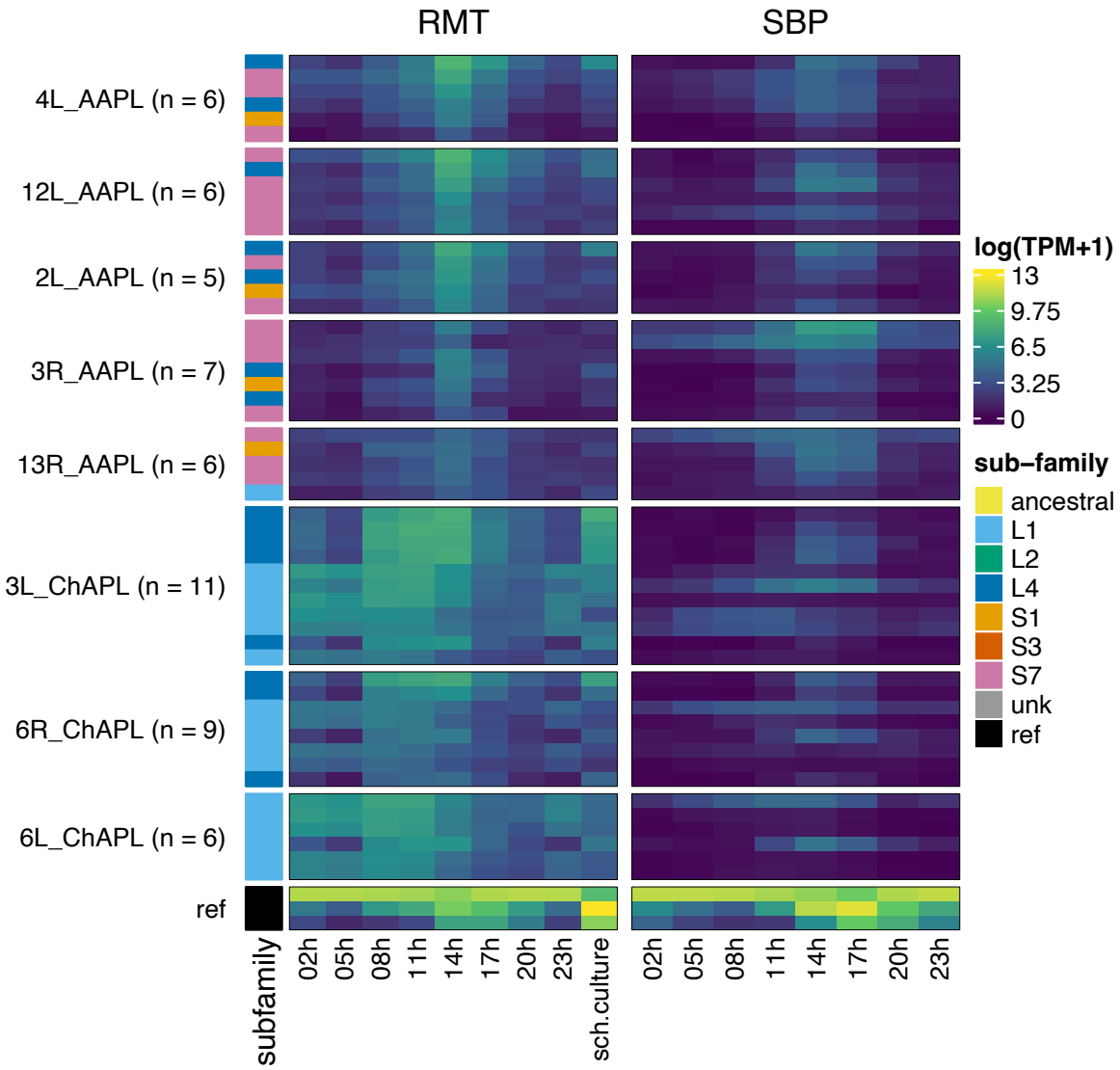


Figure 4.11 - Heatmap of the log(TPM+1) values of the ChAPL/AAPL *pir* genes across the twenty-four-hour cycle of *P.c. chabaudi* AS in both RMT and SBP transmitted infections, and RMT schizont cultures.
Genes are grouped and the heatmap split by the individual AAPL and ChAPL loci of the *pir* genes. The number and letter before the AAPL/ChAPL correspond to the chromosome number and sub-telomere location (left L or right R chromosome end) of the locus. Ref genes are three housekeeping genes for comparison; HSP70, MSP1 and AMA1. Genes were filtered to remove any that were not transcribed above 10 TPM in any time-point average. Sub-family heatmap on the left-hand side of the heatmap shows the sub-family membership of each *pir*. Genes in each section of the heatmap are ordered by average transcription across all samples.

time-point (perhaps a loss of signal due to schizont sequestration, although the L1s appear to be transcribed more broadly across the IDC).

These patterns from the natural infections are not all evident in the SBP asexual cycle; although L1 transcription still occurs across a broad timescale, there is no signal to suggest that L4s begin transcribing earlier than the 14h time-point, and their transcription is mostly gone by 20h (see Figure 4.12). This acute peak of SBP L4 transcription across only one or two time-points better matches the behaviour of the S7s. In SBP, the first statistically significant upregulation of the entire *pir* gene family occurs at 11h, later than the RMT 8h time-point. The time-point of *pir* downregulation for both transmission methods occurs at 17h with additional significant decreases for SBP by 20h, likely because the SBP schizonts have now progressed in the IDC. The SBP L1s start being upregulated as early as the 5h time-point, however their TPM values and their post-17h significant downregulation suggest that SBP does not show the same broadly high transcription of L1s as the RMT samples, except for a few individual genes. Nonetheless the L1s remain an early-transcribing group of *pir*s in both RMT and SBP. Although these are not major differences, this suggests that SBP is an inaccurate model for *pir* transcription across the IDC (and perhaps transcription of many other genes too).

Comparing contemporaneous time-points between RMT and SBP provides statistical support for the higher *pir* transcription in RMT, as observed by Spence et al, 2013 (see Figure 4.13). When broken down by groups I saw that the extent of this enrichment varied greatly within the *pir*s. The L-form *pir*s, particularly the L4s, show the highest number of statistically significant time-point comparisons as well as the greatest log fold changes. The S-form *pir*s show quite a contrasting pattern; the only times of significant enrichment in the RMT samples compared to SBP are at the 8h time-point. At the 17h time-point the S-form *pir*s (overall, not found for the S7s specifically) are actually enriched in the SBP samples, the only time-point and gene set among those included which showed this. Among the AAPLs and ChAPLs, the latter group has a greater log fold enrichment in the RMT samples, and this is significant for almost every time-point comparison and individual ChAPL locus. Across the time-points enrichment for *pir* subsets in RMT tends to be strongest at 8h and weakest at 17h (the latter of which could be due to the ‘schizont gap’ of RMT). These differences described above are summarised in the schematic of Figure 4.14.

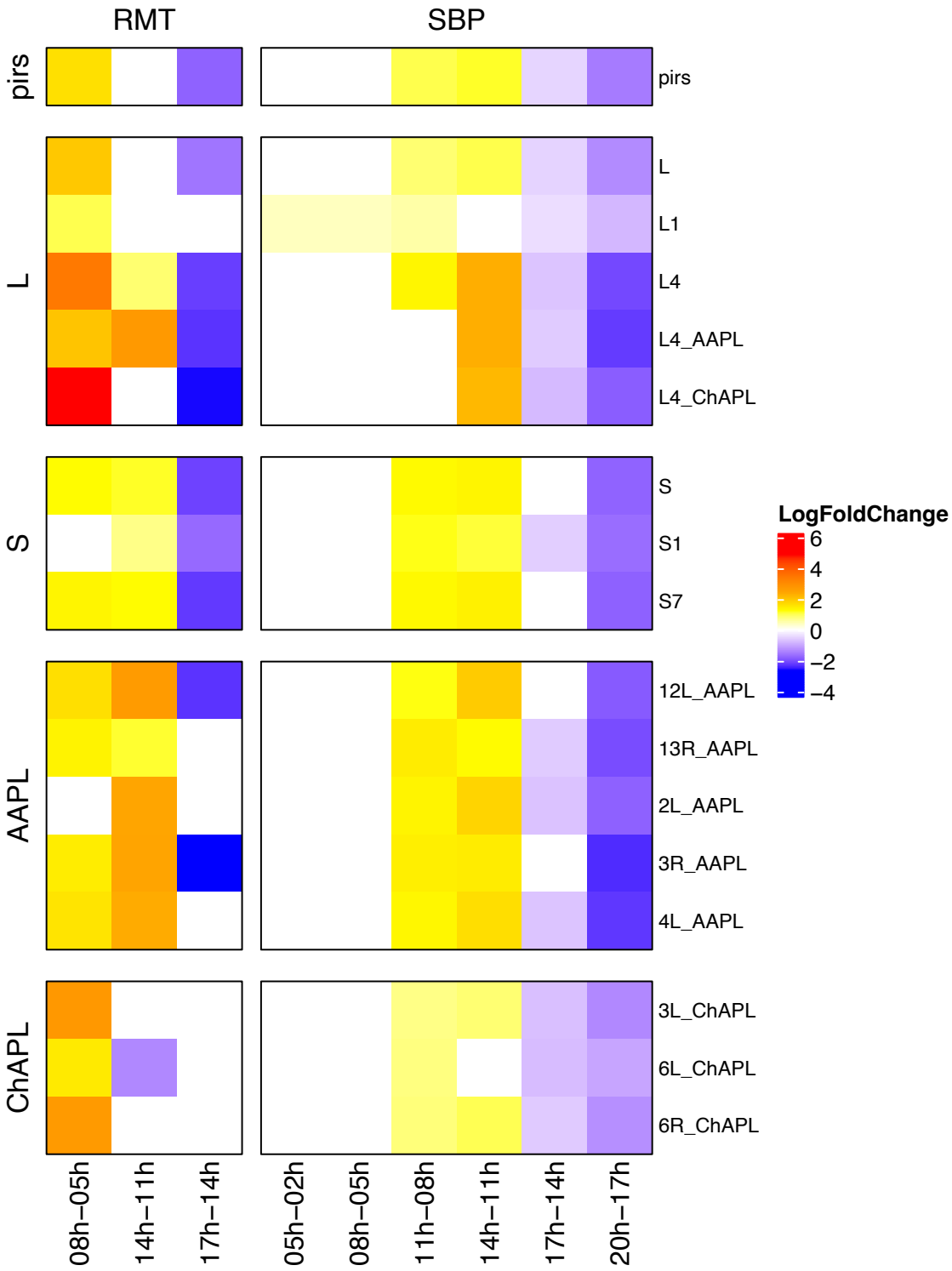


Figure 4.12 - Heatmap of the log fold change of gene set transcription comparing consecutive time-points of the asexual cycle for RMT and SBP infections.
A positive log fold change indicates higher transcription in the later time-point of the label at the bottom of the heatmap (e.g. for 14h-11h, a positive log fold change means higher transcription in the 14h samples compared to 11h). Only statistically significant (FDR <0.05) comparisons are included. Insignificant comparisons are

coloured white. The heatmap is split based on the grouping of the gene sets, including all the pirs, the sub-families and ChAPL/AAPL loci, and also split based on RMT or SBP transmission. S and L refer to all the S-form and L-form pirs respectively. L4 genes are included as one gene set, and then the L4s which are also AAPLs/ChAPLs are included are two further groups. Any gene sets with fewer than four members or no significant log fold changes are not included. Any time-point comparisons with no significant log fold changes are not included. Gene set enrichment performed using Qusage in R.

4.3.4 Common patterns of pir transcription between *P. c. chabaudi* AS and *P. berghei*

I investigated similarities between this study of *P. c. chabaudi* AS and the previous studies of *P. berghei*. While the samples presented here are only from the asexual blood cycle, there were numerous asexual blood samples analysed from *P. berghei* infections. Most of these, however, were heterogeneous for blood stages due to the asynchrony of *P. berghei* parasites (see chapter 4.1). A few exceptions include *ex vivo* generation of rings and trophozoites (from one experiment, Otto et al, 2014) and schizonts (from two experiments; Otto et al, 2014, and Modryzynska et al, 2017). With RNAseq data from these stages, I could potentially make the same analysis in *P. berghei* to contrast *pir* expression across the asexual cycle. One notable difference is that the total *pir* transcription of these stages does not appear to follow the same pattern in each species, as for *P. berghei* the progression from rings into trophozoites and then schizonts leads to non-statistically significant changes Figure 3.11. This could be a consequence of the definitions of ring and trophozoite stages used by Otto et al., 2014, as the peak of transcription in *P. c. chabaudi* AS is observed in the late trophozoite stages, just before schizont maturation and rings begin to appear in the circulation again. Hence early rings may show a high *pir* level and early trophozoites a low *pir* level if those were analysed for *P. berghei*. Although Otto et al., 2014, does not define exactly when the stages are extracted from synchronised culture, it states that the trophozoites are ‘late’ samples. Hence more synchronised asexual samples from *P. berghei* are required to verify if there is a disparity between the species. It also is not clear from these *P. berghei* studies whether the asexual blood samples were obtained from mosquito-transmitted or blood-passaged infections.

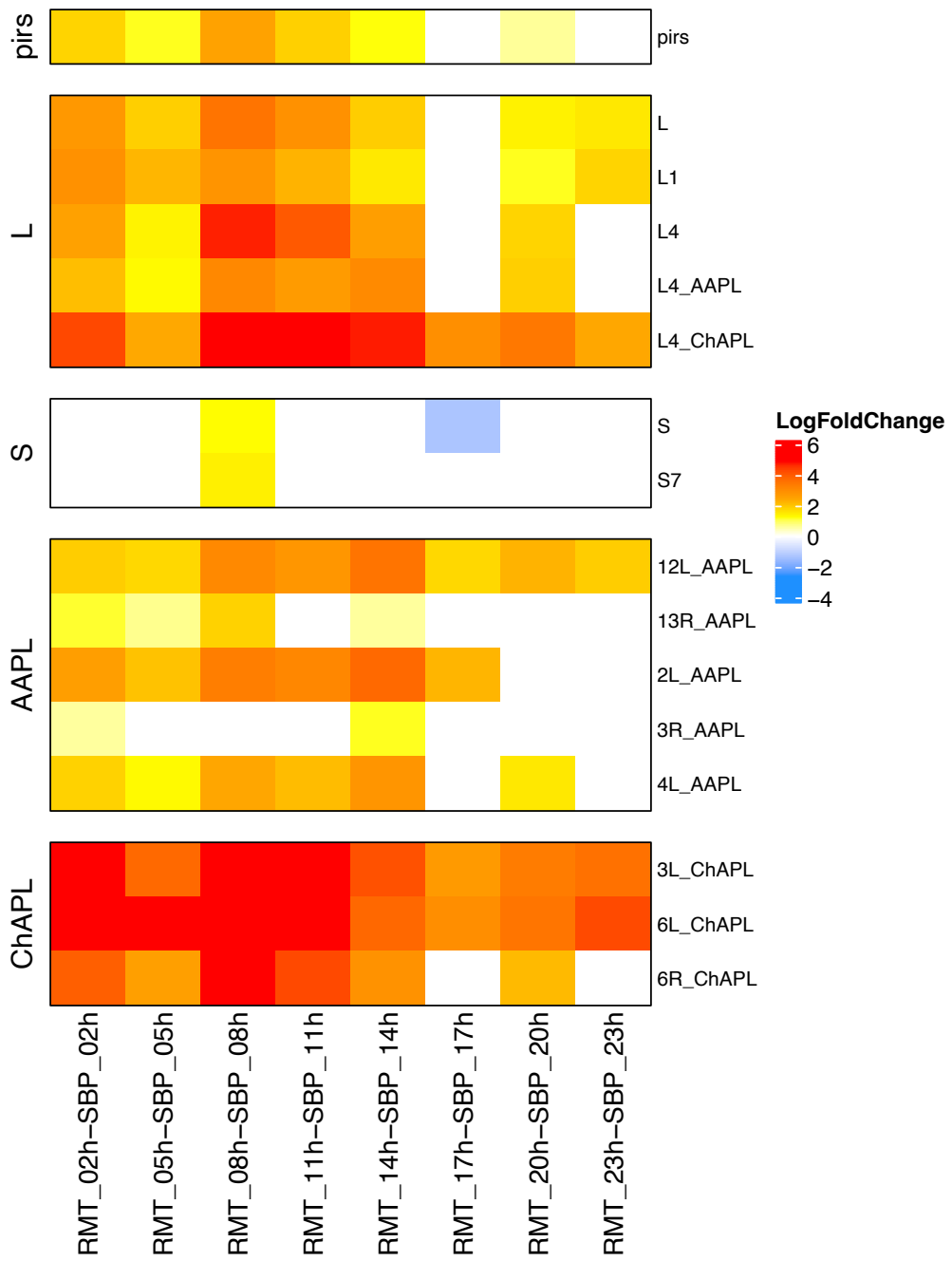


Figure 4.13 – Heatmap of the log fold change of gene set transcription comparing the same time points of RMT and SBP infections.
A positive log fold change indicates higher transcription in the RMT samples. Only statistically significant (FDR <0.05) comparisons are included. Insignificant comparisons are coloured white. The heatmap is split based on the grouping of the gene sets, including all the pirs, the sub-families and ChAPL/AAPL loci. S and L refer to all the S-form and L-form pirs respectively. Any gene sets with fewer than four members or no significant log fold changes are not included. Gene set enrichment performed using Qusage in R.

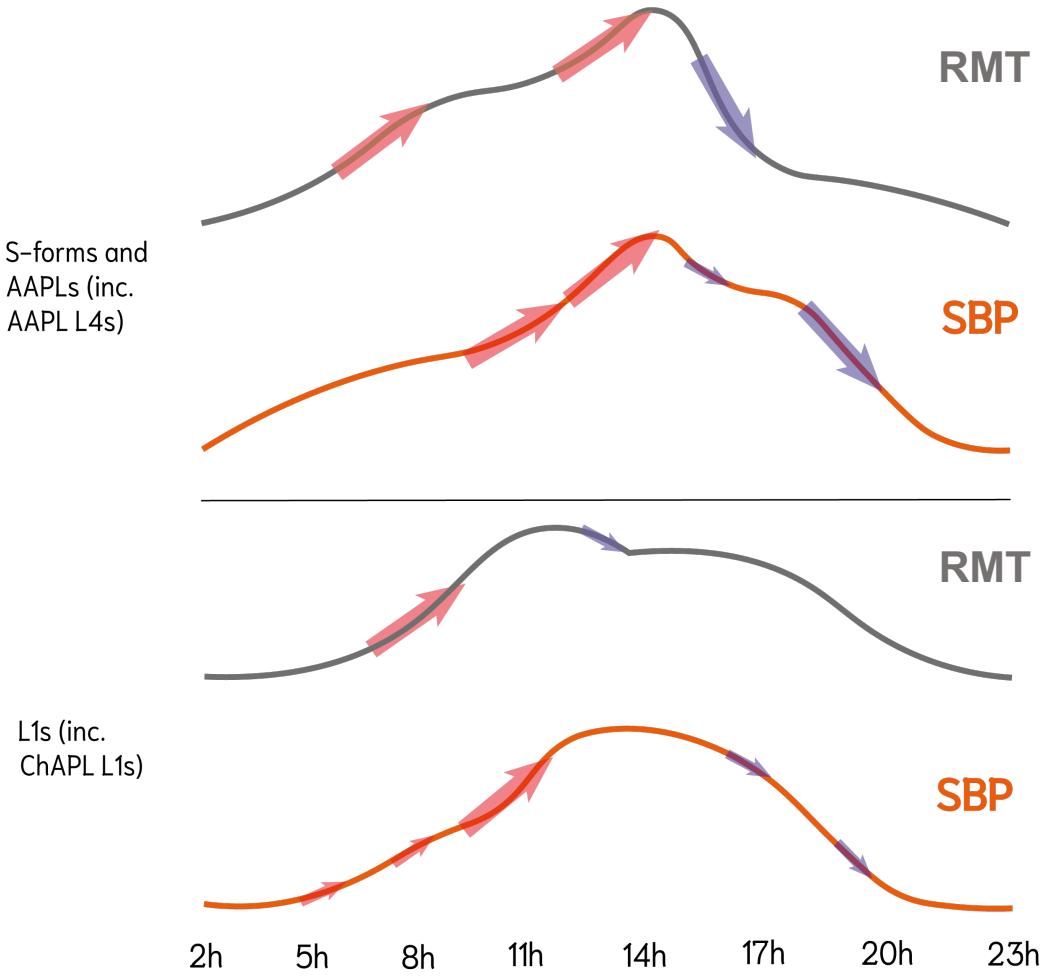


Figure 4.14 – Schematic of the approximate changes of *pir* gene transcription across the IDC for RMT (grey) and SBP (orange) parasite populations. Red arrows signify statistically significant upregulation, and blue significant downregulation, as shown in Figure 4.12. The size of the arrow roughly shows the log fold change of these transcriptional changes/whether all of the gene set or only a subset of the gene set showed the behaviour. The sharper peaking of S-forms/AAPLs/AAPL-L4s is demonstrated, while the L1s show a broader peak over time. Overall *pir* transcription is greatest at 14h.

Another difference between these *Plasmodium spp.* are the sub-family group members. Although *P. berghei* and *P. c. chabaudi* AS share Short and Long-form groupings, there are some subsets of these groups only present in the genome of one and not the other. The L3s, S2s, S4s, S5s and S8s are found in *P. berghei* and not *P. c. chabaudi* AS, and vice-versa for the L4s, S3s and S7s. Hence the only sub-families they have in common

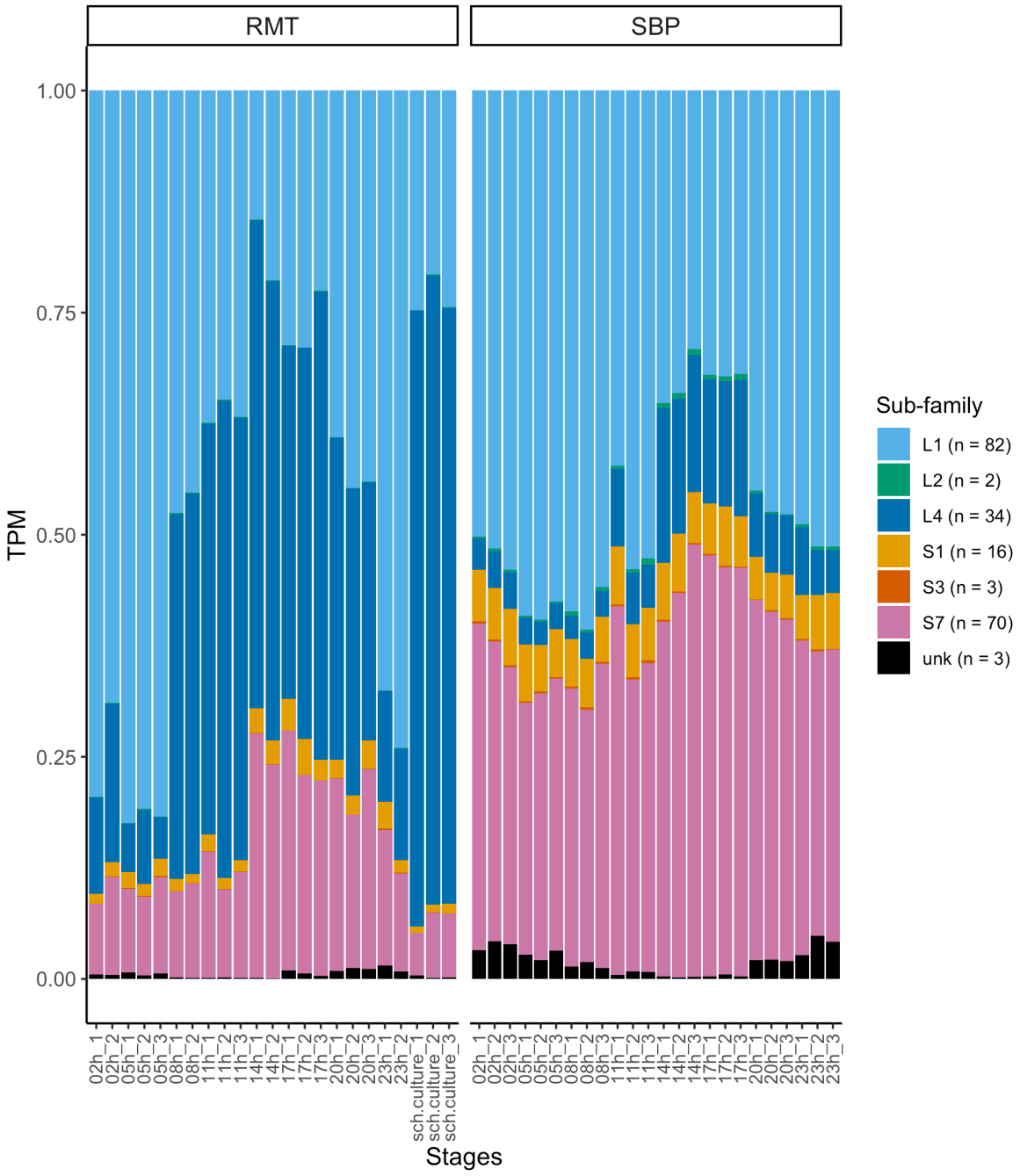


Figure 4.15 – Filled bar chart showing the proportion of *P. c. chabaudi* AS non-ancestral *pir* sub-family transcription across every sample.
Bars are coloured by sub-family, ordered by L-forms, then S-forms, then unknown clade.
The chart is facetted by RMT (left) and SBP (right) transmission methods.

are the L1s, L2s and S1s. Brugat et al., 2017, observed similarities between the *P. c. chabaudi* AS L1s and some chronicity-associated *P. berghei* L2s. The numbers of Short and Long-form *pir*s in each species are skewed in opposite directions: *P. c. chabaudi* AS

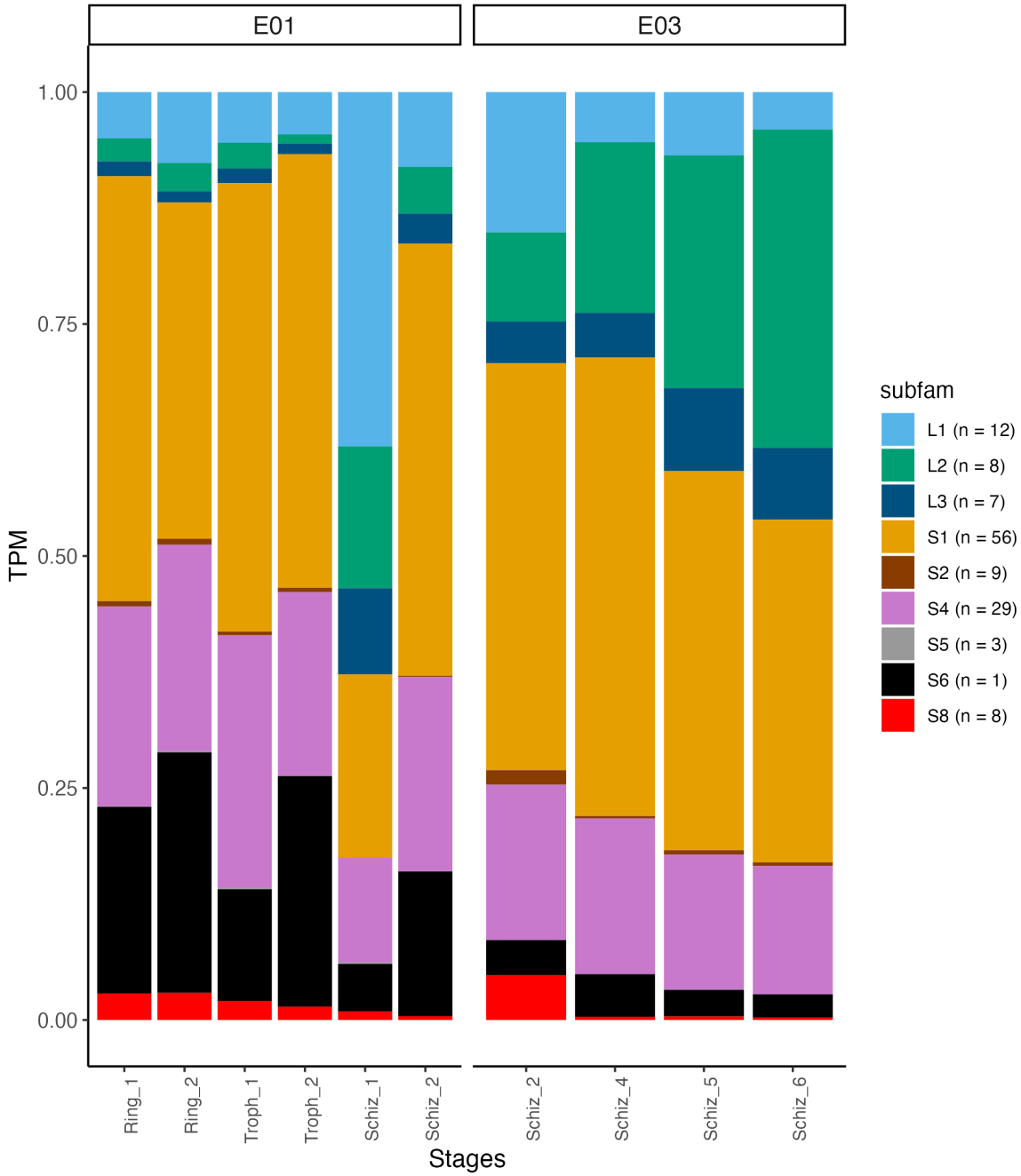


Figure 4.16 – Filled bar chart showing the proportion of *P. berghei* non-ancestral *pir* sub-family transcription across the non-mixed asexual samples of E01 and E03 (see Table 3.1); Bars are coloured by sub-family, ordered by L-forms, then S-forms. The chart is facetted by E01 (left) and E03 (right).

has 118 Long-form and 89 Short-form *pir*s, and *P. berghei* has 27 Long-form and 106 Short-form *pir*s. Therefore, it is not unexpected that infections of *P. berghei* predominantly transcribe S-forms (see Figure 3.9) and RMT infections of *P. c. chabaudi* AS mostly transcribe L-forms (Figure 4.6). The SBP infections show a much more S-

form skewed infection, however. This phenomenon is observable in Figure 4.6, the lack of significant downregulation of S-forms compared to RMT in Figure 4.12, and clearest in Figure 4.15 which shows proportions of sub-family transcription in the samples (excluding the ancestral *pir*). Driven primarily by relative S7 upregulation, there are some SBP samples for which S-form transcription constitutes the majority of *pir* sub-family transcription e.g. the 17h time-points. It is interesting that both *P. berghei* infections and SBP *P. c. chabaudi* AS infections are more virulent and express more S-form *pir*s than more benign RMT *P. c. chabaudi* AS infections.

Notably this runs counter to the paradigm that the more-virulent chronic *P. c. chabaudi* infections are dominated by L-form ChAPL transcription (Brugat et al., 2017). The most proportionally L-form enriched RNAseq samples of the *P. c. chabaudi* asexual cycle were the RMT schizont cultures, with 94-92% of non-ancestral *pir* TPM coming from L-form genes in these samples (see Figure 4.15). The schizont RNAseq samples obtained in previous experimentation on *P. berghei* suggests that this life cycle stage is similarly L-form skewed, even though L-forms are much less abundant in the genome (see Figure 4.16 and Otto et al., 2014). In RMT *P. c. chabaudi* AS RNAseq the enrichment of L-forms is mostly due to L4 *pir* transcription, while in *P. berghei* this increase of transcripts is distributed between the L1, L2 and L3 *pir*s. I performed statistical tests to compare the proportions of L- form *pir*s relative to all non-ancestral *pir* transcription, separately for *P. c. chabaudi* AS and *P. berghei*. When comparing the proportion of L-form *pir* transcription of the three RMT *P. c. chabaudi* AS schizont culture samples to the rest of the RMT asexual cycle, every pairwise comparison showed a significantly (adjusted $p < 0.05$) higher proportion in the schizont cultures except for samples 2 and 3 compared to 02h sample 1. Similarly, the only non-significantly higher L-form proportions found when performing pairwise comparisons of the E01 and E03 *P. berghei* schizont samples to the E01 ring and trophozoite samples were E01 schizont sample 2 compared to E01 rings sample 1 and trophozoite sample 1. In summary, most of the tests of whether there is higher proportionality of L-form transcription in the schizont sample versus other defined asexual cycle stages came back positive, for both *P. c. chabaudi* AS and *P. berghei*. As a final point of evidence, the statistical comparison performed for Figure 3.11 showed that when transitioning from asexual schizonts to male gametocytes, although the *pir* family is significantly upregulated as a whole, the L2 and L3 sub-families in contrast are downregulated.

4.4 Conclusions

Given that the closest links between *pir* genes and possible functions have been made from *P. c. chabaudi* AS in the host blood, gaining greater understanding of the transcriptome of this period is essential. Using RNAseq of *P. c. chabaudi* AS taken from time-points across the asexual cycle in the blood, seven days post-infection, I have been able to quantify the change of *pir* gene transcription as asexual development progresses. With data from parasites obtained from a recently mosquito-transmitted infection and from serial blood passage I have compared how the patterns change depending on level of virulence in the host and determined that the latter is an unsuitable model for this analysis. The peak of *pir* transcription occurs at 14h, suggesting that late trophozoites are the major *pir* transcribers of the *P. c. chabaudi* AS IDC and a possible role of PIRs in schizont/merozoite biology. Phylogenetic and chronicity-derived groupings of *pir* genes show similar but different patterns of transcription, peaking at 17h but with L-forms/ChAPLs exhibiting broader, earlier transcription and S-forms/AAPLs demonstrating a more sharply demarcated peak. In addition, the L-forms are statistically enriched in schizonts, corroborating observations from *P.* berghei, and pointing towards a particular conserved function of L-forms in schizogony/merozoites across rodent malarias. Although the ancestral *pir* still peaks at 14h/schizonts similar to other members of the family, it is transcribed highly across the entire cycle and remains one of the most transcribed *pir*s, reaffirming that this gene may be of particular importance for *Plasmodium spp.*.

Plasmodium spp. are thought to transcribe shortly before translation, with a delay of just a few hours in the asexual blood cycle. This ‘just-in-time’ model was predicted by RNA measurements and concomitant protein measurements (Foth et al., 2011). This means that the PIR product may be needed not for the late trophozoites but the next stage; the schizonts. Indeed, the schizonts themselves are high *pir* transcribers, and it is likely that many instances of *pir* transcription drop-off observed at the 17h time-point in RMT are due to the ‘schizont gap’ when these cells sequester. Given this timing, *pir*s may be involved in the proliferation machinery of schizogony, or perhaps a merozoite function such as cell invasion ligand-binding.

The peak of *pir* gene transcription at 14h may coincide with a timing of enhanced variability in the malaria parasite transcriptome. Comparing microarray transcriptomes of

several *Plasmodium sp.*, including *P. c. chabaudi* AS, it was found that the genes that peaked in the late trophozoite stages of *P. falciparum* were significantly enriched for orthologous genes which had shifted peaks of transcription between species (Hoo et al., 2016). Similarly, this was seen for genes which peaked in the early schizont stages for *P. knowlesi*, *P. vivax*, and *P. berghei*. *P. c. chabaudi* AS only had statistically significant enrichment for genes with inter-species shifted peaks in the early ring stages, which was seen for every species analysed. Although Hoo et al., 2016, did not see the gene peak variability at this time-point in their *P. c. chabaudi* AS samples, it is clearly a time of enhanced variability for other closely-related species *P. vivax*, and *P. berghei*. This may represent a time of the life cycle when the conservation of expressed genes is less important, such as for invasion genes, and the different species have evolved new transcriptional strategies from shared genes. Although the *pir*s (excluding the ancestral *pir*) were not included in that study as they are not orthologous across species, they peak alongside divergently transcribed genes. These genes, which likely have transcription that can be varied to suit the requirements of the parasites for their host, include those involved in fundamental processes “such as intracellular trafficking, transcription, translation, mitochondrial functions and DNA replication”. If PIR proteins were involved in these mechanisms that would be a surprise, although intracellular trafficking could involve the movement of cargo to the cell surface, where PIRs are often located.

There is a discrepancy between the microscopy and RNA-based deconvolution (Figure 4.2) at certain time-points. For instance, even though gametocytes would have been enumerated as part of the counting process, their numbers were always negligible by light microscopy. However, at 17h and 20h time-points in RMT, the deconvolution results suggest a proportion of female gametocytes are present. Merozoites are also present in the deconvolution data and not the light microscopy counts, however this stage is so physically small that enumeration is not performed by eye. The gametocyte signal may be an artefact of sexually-committed asexual cells which cannot be distinguished yet by Giemsa stain. Ultimately the major distinction between the methods is that microscopy will catch morphological changes and mRNA deconvolution will capture transcriptional changes. Where one of these attributes is altered with minimal change to the other, there will be discrepancy. An interesting method to investigate this may be to do a reverse deconvolution of the microscopy proportions, i.e. how much of the RNAseq data, and which transcripts, can be explained by the stages counted by eye? Highly

transcribed genes which are not explained by the count data may be due to stages missed by microscopy (such as merozoites or early transitions) or shifts in gene transcription not captured by the scRNAseq data which underlies the deconvolution method. Ultimately the best way to overcome any complications from stage composition would be to perform single-cell RNAseq for these infections. This would, for example, identify asexual trophozoites which are sexually committed which could be mis-identified as female gametocytes by deconvolution – possible sexual commitment signals have been found as early as liver stages (Mancio-Silva et al., 2022). ScRNAseq would also open the chronic phase IDC to analysis, which would enable comparison of more/less virulent parasites within an RMT-initiated infection. The parasitaemia of chronic *P. c. chabaudi* AS is so low that it is difficult to extract sufficient material for bulk RNAseq, especially for RMT infections which have even fewer parasites than the corresponding times in SBP. For bulk RNAseq, one would have to raise the parasitaemia of chronic infections by passaging the chronic parasites into an immunocompromised mouse, or by therapeutically depleting the immune response in immunocompetent mice during recrudescence.

When it came to the RNAseq data itself, a substantial source of variation within the RMT samples, and a major distinguisher with the SBP samples, was a single pseudogene, *PCHAS_0420800*. The differential expression of *PCHAS_0420800* was so strong that it alone dominated the second largest principal component of the data (Figure 4.3). Using the information on PlasmoDB (Aurrecoechea et al., 2009) it can be seen that this pseudogene is annotated as being the terminal gene of chromosome 4, within sub-telomeric sequence, and that it neighbours a *pir* gene *PCHAS_0420700*. The transcript length is predicted to be 981bp long, and the existence of the RNA transcript also has substantial evidence from analysis of the *P. c. chabaudi* IDC from Subudhi, et al., 2020. The RNAseq data from Spence et al., 2013, also supports the assertion that this is an MT-high pseudogene. Translation of the sequence reveals a high proportion of histidine; the largest predicted open reading frame (from the first frame in the 5’3’ direction) of 220aa is 58% histidine residues. This translation also reveals the four stop codons which likely led to the annotation of *PCHAS_0420800* as a pseudogene. For my analysis I removed *PCHAS_0420800* as it was a significant cause of unwanted variation. I, therefore, assumed that the RMT samples account for the production of *PCHAS_0420800* RNA by increasing global transcription, such that the proportions of other genes can be comparable to the SBP samples with low *PCHAS_0420800*

transcription. I am also assuming that this is not a pseudo-pseudogene: a misannotated pseudogene that actually possesses function either as an RNA molecule or a processed polypeptide. A number of examples of misannotated pseudogenes have now been described, such as a transcription-regulating RNA from an ABC transporter 'pseudogene' in humans (Piehler et al., 2008) and read-through translation of an olfactory 'pseudogene' in *Drosophila sechellia* (Prieto-Godino et al., 2016). Proteomics continues to identify 'pseudogenes' which actually have peptide products (Wright et al., 2016). Given the high levels of transcription after vector transmission, further investigation of *PCHAS_0420800* would be important to rule it out as an important player in the differences between MT and SBP.

Once again, the ancestral *pir* is a consistently highly transcribed *pir* gene. It is transcribed across the entire asexual cycle and most of the time it is the highest transcribed *pir*. In RMT there is a major drop-off in transcription at 17h, however the high transcription in schizonts and corresponding SBP time-points suggest that this is an artefact of the 'schizont gap'. Despite being transcribed constitutively, the ancestral *pir* still peaks in transcription at 14-17h like most other *pir*s and the family as a whole, suggesting that it has an important function throughout but is mostly needed in late trophozoites/early schizonts. This also indicates that its function does not diverge too far from the function of other members of the family. Although SBP parasites have diminished *pir* repertoires the ancestral *pir* still presents a strong signal, so this gene is still transcribed when other *pir*s are deprioritised.

Division of *pir*s into Short or Long-forms, or AAPLs or ChAPLs, shows that these groupings define transcriptionally distinct *pir*s. Transcription of the Long-form *pir*s is upregulated more strongly earlier in the IDC than the Short-form *pir*s. Although they are both significantly upregulated at the 8h time point (early trophozoites), the Long-form *pir*s experience a greater log-fold change and experience no further significant upregulation to reach their peak. The L1 sub-group and the ChAPL *pir*s behave similarly, which is understandable given the extensive shared membership. Similarly, the S-form and AAPL *pir*s overlap and have comparable transcriptional behaviour. One group that bridges this divide is the L4 subfamily, which contains members of both AAPL and ChAPL loci. Indeed, separating these two L4 gene sets suggested that the L4s are better defined by whether they are also members of a ChAPL or AAPL locus. It would be worth investigating whether these genes phylogenetically diverge too, to see if the

sequences have changed based on their different loci. The potential functional relevance of the ChAPLs even in RMT is surprising, given the similarities between the virulence of SBP and chronic parasites. The ChAPL transcription seen in these acute infections may be the pattern selected for in the parasites that grow in the chronic phase. ChAPLs are still more highly transcribed, and over a longer period of time, in RMT parasites, despite the fact that SBP parasites recrudesce earlier (Brugat et al., 2017). This may be a consequence of the broad transcription of *pir*s in RMT parasites, which includes the ChAPLs, whereas in later chronic infections the ChAPLs are the few remaining components of a smaller *pir* repertoire. The variation in the duration of gene transcription is an observation that could not have been made without looking across the IDC.

It is difficult to hypothesise why one group of *pir*s may be required slightly earlier in the intra-erythrocytic life cycle than another. This could potentially be investigated by targeted knock-down of the entire L1 contingent in a *P. c. chabaudi* AS isolate. Similar mass targeting screens have been achieved using CRISPR-interference: the guide RNA mapped a conserved region of a GC-rich non-coding RNA family (which regulated *var* gene transcription) (Barcons-Simon et al., 2020). Using the existing data, one could also analyse which other gene group transcriptomes correlate with the different *pir* subsets, however care needs to be taken to account for the statistical implications of TPM data being a proportion (Lovell et al., 2015; Pearson, 1897).

It is clear that using SBP parasites alone would give rise to different conclusions from those obtained from RMT parasites, as some of the patterns of RMT *pir* sub-family transcription are not clear or are totally absent in SBP. These differences demonstrate how SBP, despite being historically the standard method of maintaining experimental infections, does not capture the natural patterns of *P. c. chabaudi* AS *pir* gene transcription. From the microscopy stage enumeration and the RNA deconvolution (Figure 4.2) a caveat was identified in directly comparing RMT and SBP; the proportions of each asexual developmental stage were different between the methods of infection. The most notable difference is the presence of schizonts in the circulation of SBP infections around 17h, while their abundance is negligible in RMT. Schizonts are thought to be absent from the circulation for most strains of *P. c. chabaudi* and many *Plasmodium spp.* since they bind the vessel epithelium and do not continue flowing through in the blood (Cox et al., 1987). Cox et al., 1987 observed that a particular strain, PC7, that did not seem to sequester (in the liver), had greater schizontaemia than other

strains that did sequester. Indeed, this strain also appears to be particularly lethal to infected mice, although it is unclear from the publication whether this virulence is just due to the SBP transmission, or whether schizont sequestration is influenced by number of passages (Hommel et al., 1982). So, when schizonts are found in the circulation this is likely a consequence of a deficiency of binding the epithelium. Cox et al., 1987 also observed negligible schizont levels in the blood infections of a *P. c. chabaudi* AJ strain selected for sequestration. Although there may be a greater proportion of blood-borne schizonts in SBP infections, the parasites still sequester in the liver, lungs, and spleen after being transmitted this way, like RMT (Cunningham et al., 2017). Other studies have shown that mouse knock-outs of RAG, IFNγR and IL-10, do indeed alter levels of sequestration of *P. c. chabaudi*, meaning that immunity influences sequestration patterns (Brugat et al., 2014).

The stage composition is the easiest explanation for the 'phase shift' of some genes between after blood passage of parasites. With this dataset, it is difficult to tease apart the effect of biological difference *within* the individual stages of the life cycle, but that cannot be excluded. To resolve these differences (aside from using single-cell RNAseq) one could make use of identified rhythmic genes in *P. c. chabaudi* and observe how these vary across the RMT/SBP samples, such as those defined by Subudhi et al., 2020. Serpentine Receptor 10 is one such gene, and also a regulator of the rhythmicity of many genes. Timing of the asexual blood cycle is influenced by environmental cues (Prior et al., 2018), but in these experiments some of the possible signals, such as day-night cycles and feeding times, have been kept constant. However, RMT and SBP induce different host immune responses, including levels of immune cells, IgG and cytokines (Spence et al., 2013). Immunity can confer an alteration to the circadian cycle, as observed with the de-synchronicity of *P. c. chabaudi* AS when infecting $RAG^{-/-}$ and $TNF\alpha^{-/-}$ mice (although mice with disrupted T-cells or immunoglobulin responses did not show a loss of circadian regulation) (Hirako et al., 2018). Preliminary results in *P. berghei* also suggest that lipopolysaccharide-induced inflammation delays progression of the IDC (Lansink et al., 2021). Variations in the peak period calculated from rhythmic genes that match the overall transcriptional 'shift' may suggest that the differences observed are due a life cycle shift and not gene transcription changes at the same point of the life cycle. Alternatively, mixture models could be applied to tease out the impact of the developmental timing sub-populations.

The S-form *pir*s are the only ones to show statistically significant downregulation at any timepoint when comparing time-matched RMT and SBP samples, at 17h. This could be due to the RMT schizont loss in the RNAseq data mentioned above, however the heatmaps do not suggest that S-form transcription in the RMT schizont cultures is particularly high (Figure 4.10). As well as the changes in the amount of transcription of these groups, there are also distinctive patterns in when the sub-families/loci start being transcribed in the cycle, and for how long. The statistical evidence for the different times of up- and down-regulation of these gene sets within both transmission methods demonstrates that there is an intrinsic change in gene transcription timing and the differences are not just a consequence of the smaller overall *pir* repertoire in SBP. The shift in SBP parasite populations towards less overall *pir* transcription but with a higher proportion of S-form transcription matches the model of a more 'refined' *pir* transcriptome for these mouse-adapted infections.

As previously reported, a broad swathe of the *pir*s are transcribed more highly in RMT samples than time-matched SBP, however this pattern is not the same across the entire IDC. The previous major comparison by Spence et al., 2013, used a singular time-point at 8h (in accordance with the times used in this thesis). Using the data provided here would give the same conclusion, as shown by the differential expression of major gene sets of the *pir*s (Figure 4.13). Indeed, this is one of the times when the difference between *pir* transcription of RMT and SBP is particularly high. If singular time-points were taken at other points of the IDC, the outcome could have been different. A comparison at 17h with these data may have led to a different overall conclusion, as the *pir* gene set would not have enough evidence to be deemed differentially expressed. The SBP-high *pir* defined by Spence et al. only appears to show the expected behaviour at 17h in this experiment (Figure 4.8) – and this too may be an artefact of schizont loss in RMT. At 8h, when Spence at al. took their samples, schizont differences should not be impacting the data as both their results and Figure 4.2 indicate that schizont proportions in SBP are low and trophozoites dominate (Spence et al. reported average >97% trophozoites in both conditions). Given that SBP infections preferentially transcribe a smaller subset of *pir* genes, it may be that this subset changes and, for these experiments, this subset does not happen to include the SBP-high *pir*. Note that the SBP-high *pir* is still transcribed highly, but there is no major difference between SBP and RMT infections. There is no other obvious candidate for an SBP-high *pir* among the data included in this thesis, although the S-form *pir*s do form a subset of the multigene

family which the SBP parasites sometimes transcribe more highly than the RMT population. These observations highlight the importance of covering multiple time-points across the IDC when investigating *Plasmodium spp.* intraerythrocytic biology, as well as using more biologically-relevant mosquito transmitted infections.

Comparison of *pir* transcription between species is not trivial because of the different sub-family constitution of *pir*s in each genome. *P. berghei* and *P. c. chabaudi* are probably the closest models for which extensive RNAseq data exist, making them the best for comparison. Even though *P. berghei* has an S-form dominated genome, and *P. c. chabaudi* AS the opposite, the *P. berghei* and *P. c. chabaudi* schizont samples both showed an L-form enrichment. This is far from a conclusive result as the *P. berghei* schizont sample *pir* repertoires in Otto et al., 2014, are not that similar to one another and it would have been more desirable to have additional replicates. The Modryzynska et al., 2017, samples have more replicates and show better consistency (among the samples included after quality control); however that experiment did not extract any other synchronised stage of the IDC for direct intra-experimental comparison. Given the frequent issue of asynchrony in *P. berghei* infections, and sequestration in *P. c. chabaudi* AS, single-cell RNAseq may be the best method to investigate L-form transcription in schizonts in the future, even if this limits analysis to just the dominating *pir*s. Male gametocytes of *P. c. chabaudi* AS should also be included in future analysis, as they appear to have strong L-form (particularly L4) enrichment also, which is not seen with *P. berghei* male gametocytes (Cunningham and Reid et al., in review).

The evidence from *P. berghei* does not suggest a similar overall pattern of total *pir* transcription in the asexual blood stages (see Figure 3.11) as transcription shows no statistically significant change between ring and trophozoite stages. This could be an artefact of needing to synchronise *P. berghei* and then grow it *in vitro* to delineate different stages of the IDC, as well as the cellular stress response external culture may induce in the parasites. *P. vivax* microarray data showed that the *pir*s tended to be part of a group upregulated in the early-late ring stages, although there was also a smaller group of late trophozoite/early schizont *pir*s which corresponds better with these data (Bozdech et al., 2008). This smaller microarray data set was corroborated by RNAseq data based on the same samples, which is re-analysed and discussed in depth in Chapter 5 (Zhu et al., 2016). *P. knowlesi* microarray transcriptomes suggest that the largest significant upregulation of *pir*s occurs at the early trophozoite stage, although this

is based on four genes out of only eight found to be differentially expressed (Lapp et al., 2015). Evidence from other *P. berghei* stages however shows that high *pir* transcription occurs at times when the parasite is preparing for proliferation, such as late liver schizonts and male gametocytes. The most highly transcribing samples found by my analyses, the liver merozoites, may suggest that the PIRs are important for an aspect of highly-motile short-lived stages of parasite development. Indeed, the timing of *pir* transcription coincides with that of DNA replication and deoxyribonucleotide synthesis in the *Plasmodium falciparum* IDC, in anticipation of proliferation (Bozdech et al., 2003). These are indications that *pir*s are transcribed for times when the parasite is mass producing daughter cells.

Chapter 5. Results - Analysing the *P. vivax pir* transcriptome using *de novo* transcriptomics

5.1 Introduction

Plasmodium vivax is the major human malaria parasite after *P. falciparum*. While *P. falciparum* is most prevalent in Africa, *P. vivax* is the major contributor to malaria infections in the Americas and most of Asia. Globally the proportion of all malaria cases attributable to *P. vivax* has been decreasing since 2000, however in South-East Asia this figure has remained at about 54-35% over the last two decades, while in the WHO Western Pacific region it has increased from around a fifth to a third of all human malaria infections. The parasite is prevalent in the Americas, where 67-78% of malaria infections are accountable to *P. vivax*, peaking in 2018 at 78.2% (WHO, 2021). It has traditionally been seen as unable to penetrate into the African continent to any great degree due to the high Duffy antigen-negativity among people (Boyd & Stratman-thomas, 1933; Hamblin & Di Rienzo, 2000; Mayne, 1932; Miller et al., 1976). However, it has since been shown that *P. vivax* can indeed invade erythrocytes independently of the Duffy antigen and cause infection in Duffy-negative individuals (Ménard et al., 2010). Increasing evidence of seropositivity in African nations, and significant sequestration of parasite in the spleen (away from the circulation), have led some to suggest that *P. vivax* in the continent is much more endemic and problematic than common mantra suggests (Baird, 2022; Culleton et al., 2009; Kho et al., 2021). As a consequence, *P. vivax* presents an unknown threat to African populations. Given its prevalence around the world and the possibility of hidden transmission/future intrusions into Africa, *P. vivax* is a major focus of public health research.

P. vivax and closely related species exhibit a number of biological distinctions to *P. falciparum* of importance to human infection. *P. vivax* has a strong preference for immature erythrocytes (reticulocytes) and possesses a quiescent liver stage of development, the hypnozoite, which can effectively hibernate, escape antimalarial treatment, and cause recrudescent infections in hosts (Kitchen, 1938; Krotoski, 1985). Additionally, *P. vivax* is the major human malaria parasite to have *pir* genes; indeed, *pirs* were originally discovered in this parasite (del Portillo et al., 2001). *P. vivax* contains 1216 members of the *pir* gene family in the P01 strain reference genome, constituting

around a sixth of the total number of predicted genes for this strain (Auburn et al., 2016). A recent high quality genome assembly of the Thai-origin W1 contains as many as 1145 predicted *pir* genes (Minassian et al., 2021). This suggests that the comparatively few *pir*s found in earlier genome assemblies, such as Sal1 (Carlton et al., 2008), were artefacts of the limited technology of the time as opposed to biological differences. The major dedication of a human malaria genome to a single multigene family means that greater understanding of its role in infection is of high importance.

While rodent malaria parasites have been used extensively for *pir* investigations, as detailed in previous sections, these models cannot infer the true variation within the gene family, as these parasites have been highly selected after many years of culturing in the laboratory. This is an advantage for minimising the variability of results between experiments, as genetic variation is low and reference genomes are close matches to the specimen under study. However, they probably do not reflect the extensive variation of the *pir* multigene family in wild isolates, and thus their use for elucidating *pir* functions across species, including human plasmodia, could be limited. Among the different species, *P. vivax* is particularly diverse between isolates, and this is heavily concentrated within multigene families such as the *pir*s (Neafsey et al., 2012). Sequenced genomes across different geographic regions show that there are differences not just in the *pir* repertoire but also in the proportions of *pir* sub-family members (Merino et al., 2006; Chen et al., 2017). In transcriptomic studies, *pir*s may be missed when mapping divergent clinical isolates to a generic, probably unrepresentative, reference genome. Therefore, studies that seek to answer key questions about the *pir* gene involvement in *P. vivax* infection, such as when they are transcribed and which sub-families are transcribed at different stages, have major limitations if they use the currently annotated genomes.

One approach to obtain a more comprehensive overview of the number and variability of pir genes in *P. vivax* is to use *de novo* transcriptomics on the available *P. vivax* RNAseq data, a method in which the sequenced reads are used for assembly of the RNA being transcribed, without need of a reference genome. In this process, the reference is constructed from the RNAseq output itself. Hypothetically this technique could identify unknown *pirs* in data from disparate *P. vivax* isolates. Similar analyses conducted for the *var* genes of *P. falciparum* have revealed high rates of variation in novel transcripts between individual isolates. One example is a study that used *de novo* transcriptomics

to elucidate *var* genes being transcribed in clinical samples, which was then used to guide peptide-identification in concomitant proteomics (Kamaliddin et al., 2019).

De novo transcriptomic assembly has previously been performed for species with missing or inadequate reference genomes, including the assembly of *pir* gene transcripts. An example from rodent-infecting malaria was an investigation into *P. yoelii nigeriensis* transcription, which showed that this strain possessed novel *pirs* not found in the reference strains of the *P. yoelii* reference genomes (Zhang et al., 2021). Similar methods have been applied to *P. ovale* and *P. vivax*, although most studies of the latter have not used *de novo* assembly methods (Brashear et al., 2019; Kim et al., 2017; Zhu et al., 2016). In one of the *P. vivax* studies 179 novel transcripts were annotated as *pir*-like, many with RNAseq read evidence of active transcription, providing high confidence in the accuracy of the assemblage (Zhu et al., 2016). All these studies demonstrated marked difference between the *pir* repertoires of the reference genomes and the *de novo* transcriptomes, suggesting that without the new assembly a subset of *pir*s would have been completely missed. Using *de novo* transcriptomics permits the identification of novel members of highly variable multigene families and assessment of their level of transcription.

As described above, *de novo* assembly has only been performed in a small number of *P. vivax* investigations so far. These have demonstrated the information that can be gleaned from the technique, but not gone as far as matching these *de novo* transcripts to the defined sub-families. So far, Trinity has been the only programme used for this purpose, and other tools were not reported to have been tested. I seek to compare different assemblers for the reconstruction of accurate and complete *pir* transcripts using a gold-standard model of *P. c. chabaudi* AS. In this chapter I have analysed transcription of *pir*s by *de novo* transcriptome assembly using published datasets of *P. vivax* RNAseq (see Table 5.3). This will allow me to study the kinetics and magnitude of *P. vivax pir* gene expression during the infection cycle and determine, for the first time from *de novo* transcript data, which sub-families are transcribed at particular times in the parasite life cycle and how sub-family transcription differs between isolates. I will test whether high transcription in asexual blood-stage parasites and low transcription in mosquito-stages is also true for *P. vivax.* This would lend weight to the idea that the *pir* family has conserved function between divergent species, thus rendering the rodent model infections more relevant for understanding human infection. These timings of the family

and sub-families may give insight into functional roles of the *pir*s and whether this is conserved across the Plasmodia. Hypnozoites of *P. vivax* are of particular interest to malariologists because they may facilitate recurrence of the parasite even through drug treatment (Baird, 2009; Krotoski, 1985). This dormant liver hypnozoite stage found has not been analysed for *pir* expression in this manner, and cannot be elucidated using *Plasmodium* that infect rodents, as they lack hypnozoites. I will examine whether *pirs* are transcribed in this quiescent cell type, as this may suggest that they play a role in hypnozoite biology.

5.2 Aims

- Use a well-established reference genome to test the best programme to assemble *pir* mRNAs *de novo* from RNAseq reads (*P. c. chabaudi* AS)
- Apply these methods to *P. vivax* data to construct novel *pir*s and derive the sub-families they belong to from clustering similarity networks.
- Map each RNAseq dataset to its assembled transcriptome to measure the transcription levels of the *pir*s and *pir* sub-families.
- Assess which *pir* sub-families are transcribed consistently across experiments at given stages of the *P. vivax* life cycle.
- Determine whether there is any comparable similarity between transcription patterns of *P. vivax pir*s and the *P. berghei* and *P. c. chabaudi* AS *pir*s described in chapters 3 and 4.

5.3 Results

5.3.1 *De novo* transcriptomes of *P. c. chabaudi* AS data recover known *pir*s

To demonstrate that *de novo* transcriptomic analysis will work with the *pir* genes, and to identify which software works best for my requirements, I started by performing this method on a species for which we had ample data and a well-annotated reference genome. *Plasmodium chabaudi chabaudi* AS is ideal for this as I had access to multiple RNAseq data sets of a high *pir*-transcribing stage of development (from the recently mosquito transmitted/serially blood passaged (rMT/SBP) intra-erythrocytic development cycle experiments of Chapter 4). As this is a lab isolate from the strain (AS) of *P. c. chabaudi* used for generating the reference genome, the data should deviate minimally from the genome of the sequenced parasites. As an initial proof-of-concept, all the data from the *P. c. chabaudi* AS experiment in chapter 4 were concatenated together (401,066,418 reads in total) and put through one of three *de novo* transcriptome programs; Trinity (Haas et al., 2013), RNAseq SPAdes (Bushmanova et al., 2019), and single-cell SPAdes (Bankevich et al., 2012). This is not exactly what will be done with the *P. vivax* samples as in that case I will be, as far as possible, keeping different isolates separate to prevent the mis-assembly of transcripts from different transcriptomes. Hence the number of reads for each *P. vivax* assembly will be substantially smaller than the total included for this *P. c. chabaudi* AS analysis. The purpose of this validation step using *P. c. chabaudi* AS transcripts is to ascertain whether these methods can recover high-quality *pir* transcripts at all, and which programme may work best. Hence, to minimise loss of reads for constructing transcripts, only a small amount of filtering and quality control was performed. The resulting assemblies were then aligned to the known *pir* transcripts with BLAST to ascertain the amount and accuracy of annotated *pir* recovery (see Table 5.1). As a conservative metric of performance, I have counted the number of known *pir* transcripts with a BLAST alignment to at least one *de novo* transcript in each assembly, in which the alignment covered ≥95% of the known transcript sequence with an identity of ≥95% (cov95-id95). As a less stringent measure I have also quantified the number of known transcripts with ≥75% of sequence covered. Some *de novo* transcripts cover most of the known transcript they align with, but the known transcript covers proportionally much less of the *de novo* transcript sequence, suggesting potential chimeric transcripts. Hence, I have included how many of the cov95-id95 alignments cover less then 50% of the *de novo*

Assembly programme	Total num. of transcripts assembled	cov95-id95 (percentage of all 211 *pir*s)	Median mismatches of cov95-id95 alignments	Median gap number of cov95-id95 alignments	Num of possible chimeras (percentage of cov95-id95 *pir*s)	cov75-id95 (percentage of all 211 *pir*s)
Trinity	118,514	13 (6.2%)	2.5	0.5	2 (15.3%)	26 (12.3%)
Rna-SPAdes	100,080	71 (33.6%)	0	0	23 (32.4%)	131 (62.1%)
Sc-SPAdes	108,129	79 (37.4%)	0	0	27 (34.2%)	139 (65.9%)

Table 5.1 – *De novo* transcriptome assembly statistics for the *pir* transcripts from the rMT and SBP data combined, using different assembly programs. 'cov95-id95' stands for the number of known *pir* transcripts with a BLAST alignment to at least one *de novo* transcript, in which the alignment covered ≥95% of the known transcript sequence with an identity of ≥95%. 'cov75-id95' is the same except covering ≥75% of sequence. 'Num of possible chimeras' is the number of cov95-id95 alignments which cover ≤50% of the de novo transcript – note that this may be an underestimate as shorter chimeric transcripts may not reach the 50% threshold.

transcript. We would not expect to assemble the entire *pir* transcriptome from these samples as not all *pir*s are transcribed. Table 5.1 shows that the number of *pir* genes recovered varies substantially with each programme. Trinity, one of the most commonly used *de novo* RNAseq programmes, assembled a greater number of transcripts (as expected given that SPAdes combines similar transcripts), but few of these matched the known *pir*s. The SPAdes algorithms performed well, recovering over a third of known *pir*s with 95% converage and identity, and overall these alignments were highly accurate (low median mismatch and gap number). The main possible weakness of the SPAdes assemblies was the length of the *de novo* transcripts often proving much longer than the annotated ones. This high potential chimera rate is not necessarily an error, as the transcripts may be mis-annotated. However, I will assume that the initial annotation is correct and view high possible chimera rates as a limitation of the tools. Interestingly the single-cell algorithm from SPAdes performed better for assembling *pir*s than the

programme specific for RNA. Other investigators have also found that sc-SPAdes can outperform rna-SPAdes for some metrics of *de novo* transcriptome assemblies (Hölzer & Marz, 2019). One aspect in which Trinity performed better was in the assembly of the ancestral *pir (PCHAS_0101200)*. Neither of the SPAdes methods assembled a *PCHAS_0101200* transcript which covered at least 95% of the known sequence (they do so with at least 75% sequence coverage, and 100-99.9% identity). This is an unexpected result given that the ancestral *pir* should be highly represented in the RNAseq reads compared to other *pir*s, and it could show that Trinity performed better for transcripts transcribed at higher levels and SPAdes performed better at handling low transcription transcripts (such as most of the *pir*s). This initial analysis demonstrated that full-length, high quality *pir* transcripts can be assembled from *P. c. chabaudi* AS RNAseq data, and that SPAdes potentially functions best for this process. There are also warnings about the number of longer-than-expected/chimeric transcripts appearing.

Next, I decided to pursue a test of the methods on a more realistic data set compared to the size of the *P. vivax* data. For this, I took the RNAseq data of a sample containing highly transcribed *pirs* according to Chapter 4: a 14h rMT sample. I processed this in the same manner as I plan for the *P. vivax* data to minimise contamination with unwanted reads (see the dotted outline boxes in Figure 5.1 for the workflow of these data). After fastp processing, alignment to remove mouse-derived reads, and rRNA depletion, this leaves 3,895,080 paired-end reads in total for assembly. From Table 5.2 and Figure 5.2 it can be seen that the previous results are recapitulated in the smaller dataset. Trinity assembles the highest number of transcripts but the fewest *pir*s, and the *pir*s it does assemble have the poorest accuracy. SPAdes performs better, with sc-SPAdes now clearly assembling the greatest number of *pir* transcripts (28 assemblages meet my highest criterion of similarity and more are recovered at lower stringency - see Figure 5.2), with the highest overall accuracy and lowest number of possible chimeras. The ancestral *pir* is now also accurately assembled by all three algorithms (over 95% coverage and 100% identity), demonstrating that SPAdes may have been struggling with noise when too many reads are present. Despite only having less than 1% of the number of reads accessible for transcriptome assembly than in the previous analysis shown in Table 1, all programmes produced over 10% of the former number of transcripts and experienced at most a ~78% reduction in *pir* recovery. Trinity found only three fewer *pir* transcripts compared to the first test, when it was using every available

Assembly programme of 14h rMT sample	num_transcripts	cov95-id95	Num of possible chimeras	med_gaps	med_mismatches	cov75-id95
rnas-SPAdes	12860	**16 (7.58%)**	3 (1.42%)	0	1.5	29 (13.74%)
rna-SPAdes _0.90cdhitest	12314	**16 (7.58%)**	3 (1.42%)	0	1.5	29 (13.74%)
rna-SPAdes _0.95cdhitest	12392	**16 (7.58%)**	3 (1.42%)	0	1.5	29 (13.74%)
rna-SPAdes _0.99cdhitest	12452	**16 (7.58%)**	3 (1.42%)	0	1.5	29 (13.74%)
sc-SPAdes	10387	**28 (13.27%)**	0 (0%)	0	1	44 (20.85%)
sc-SPAdes _0.95cdhitest	10209	**28 (13.27%)**	0 (0%)	0	1	44 (20.85%)
sc-SPAdes _0.99cdhitest	10330	**28 (13.27%)**	0 (0%)	0	1	44 (20.85%)
sc-SPAdes _0.90cdhitest	10143	**28 (13.27%)**	0 (0%)	0	1	44 (20.85%)
Trinity	13107	**10 (4.74%)**	3 (1.42%)	0	6.5	18 (8.53%)
Trinity_0.90cdhitest	11777	**9 (4.27%)**	3 (1.42%)	0	3	16 (7.58%)
Trinity_0.95cdhitest	12048	**9 (4.27%)**	3 (1.42%)	0	4	17 (8.06%)
Trinity_0.99cdhitest	12395	**10 (4.74%)**	3 (1.42%)	0	6.5	18 (8.53%)

Table 5.2 - *De novo* transcriptome assembly statistics for the *P. c. chabaudi* AS *pir* transcripts from a 14h rMT sample, using different assembly programs and CD-hit-est redundancy removal thresholds (90%, 95% and 99%).
'cov95-id95' stands for the number of known *pir* transcripts with a BLAST alignment to at least one *de novo* transcript, in which the alignment covered ≥95% of the known transcript sequence with an identity of ≥95%. This is highlighted in bold as it is probably the most important statistic in the table for my study. 'cov75-id95' is the same except covering ≥75% of sequence. 'Num of possible chimeras' is the number of cov95-id95 alignments which cover ≤50% of the de novo transcript – note that this may be an underestimate as shorter chimeric transcripts may not reach the 50% threshold.

read for assembly. There was a cost in accuracy for the transcripts that were assembled, but the proportion of possible chimeras fell. Of the 28 sc-SPAdes *pir* transcripts which had BLAST alignments covering ≥95% of the known transcript sequence with ≥95% identity, the alignments also covered >50% of the *de novo* sequence. Hence, fewer information/reads provided to the assemblies do not linearly lead to a reduction in the accuracy and completeness of the *de novo* transcriptome. The assemblers still function with far fewer sequences akin to many of the smaller available *P. vivax* datasets.

De novo transcriptome assemblies frequently produce many highly similar transcripts. Although these may be biological correct, due to small genetic variations like SNPs or isoforms, they clog up downstream analysis and may not be helpful for answering biological questions. Hence, the first step of post-assembly processing that I performed is clustering the sequences to remove redundant sequences using CD-HIT (Figure 5.1) (Fu et al., 2012; W. Li & Godzik, 2006). CD-HIT clusters the sequences based on identity, identifies a representative for each cluster to include, and removes the other sequences from the clusters. To test how this affects the *pir* recovery (transcripts of which may be highly similar to one another) I used CD-HIT with 99%, 95%, and 90% identity thresholds (for similarity-based removal of sequences) on these three *P. c. chabaudi* AS assemblies (Table 5.2). As expected, given that SPAdes had a redundancy removal step built-in to the algorithm, Trinity had the most transcripts removed by CD-HIT. Although none of the assemblies had any *pir*s removed by a lenient 99% threshold,

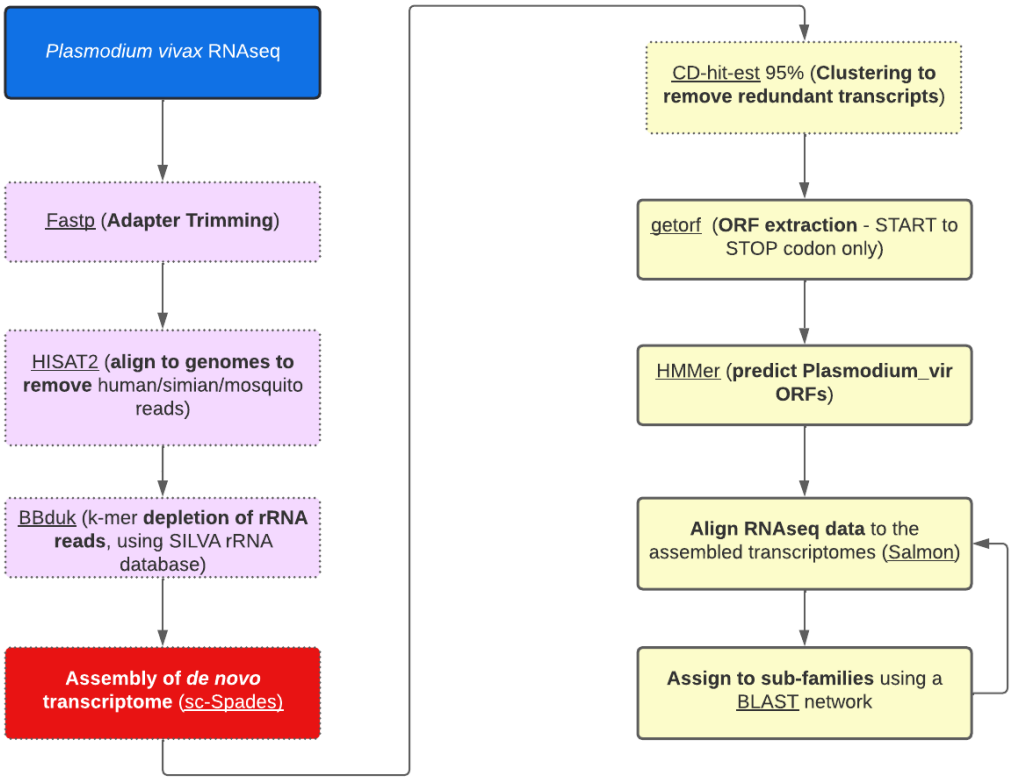


Figure 5.1 – Flow chart of the analysis of *P. vivax* RNAseq datasets to produce and investigate *de novo* transcriptomes of the *pir* genes.
The names of programs used are underlined, and the action they are performing is in bold. The pink background refers to quality control/filtering steps of the RNAseq reads, red for the *de novo* assembly itself, and yellow for post-assembly processing/analysis. Boxes enclosed by dotted lines were also performed when testing the methodology with *P. c. chabaudi* AS data (performing HISAT2 alignment to the mouse genome). Designed using Lucid Chart (lucid.app/lucidchart).

Trinity did 'lose' one *pir* at 95% and 90% thresholds. The SPAdes assemblies were unaffected. Therefore, the *de novo* transcriptome assemblers, especially those from the SPAdes family, can build accurate, distinct transcripts of *pir* genes from *P. c. chabaudi* RNAseq reads derived from individual samples.

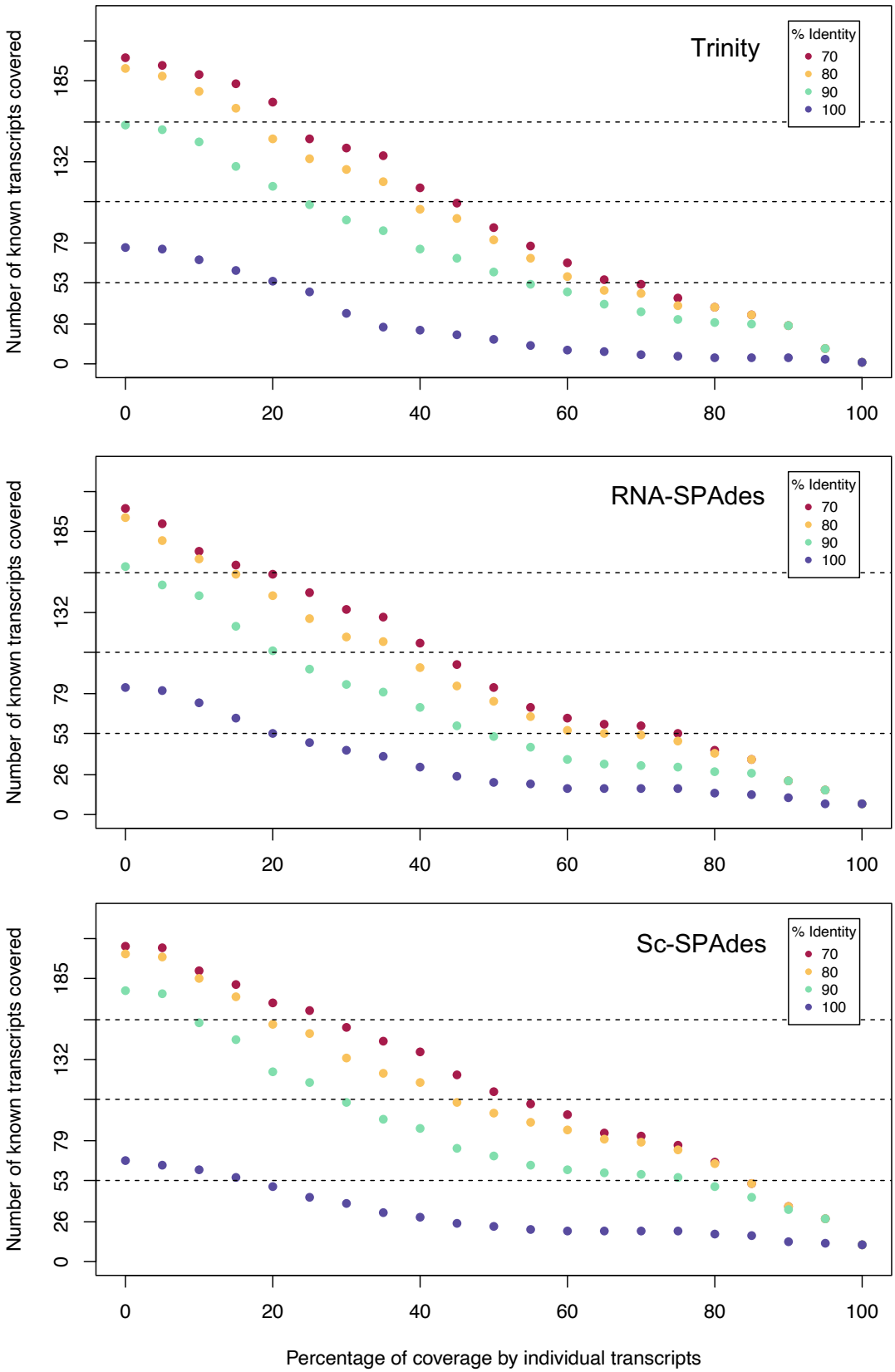


Figure 5.2 – The numbers of known *P .c chabaudi* AS *pir*s covered by *de novo* transcripts from the assemblies of a 14h rMT RNAseq sample.

The graphs are separated by the different assembly programme used, and 95% CD-hit-est was used for all assemblies. The x axis shows the threshold of percentage coverage of the known *pir* transcripts by the individual alignment/*de novo* transcript with the highest coverage. More *pir*s are found as this becomes more lenient (lower threshold of coverage). Different colours show different percentage identity thresholds for the BLAST alignments. Dotted horizontal lines correspond approximately to one quarter, one half, and three quarters of the total *pir* number (211).

5.3.2 Asexual blood cycle RNAseq of *P. vivax* can produce many *de novo* assembled *pir* transcripts

Having shown that *de novo* transcriptomics can assemble high-quality *pir* transcripts from RNAseq data, I decided to proceed with applying the best performing assembler, sc-SPAdes, to the *P. vivax* data. These data come from a variety of isolates, geographic locations, and stages of the life cycle, as summarised in Table 5.3. I chose 95% as the identity threshold for CD-HIT to remove duplicates but minimise the chance of losing *pir* information. In order to find the putative *pir* transcripts I must also identify the ORFs among these assemblages, apply a hidden Markov model to predict which ORFs have a *P. vivax pir* domain, and extract the transcripts with *pir* ORFs at least 200 amino acids long (see Materials and Methods and Figure 5.1). This workflow found 4485 transcripts with putative *pir* ORFs across all the experimental data.

The quality of these assemblies was assessed using Benchmarking Universal Single-Copy Orthologue (BUSCO) analysis to quantify how many core genes are constructed within the assemblies, comparing to both a collection of Apicomplexan BUSCOs and *Plasmodium* BUSCOs (see Figure 5.3). The BUSCO score shows the percentage of expected, conserved genes from similar organisms that were found in the assembly. In general, the assemblies with more *pir* genes predicted are in the upper right of the graph, indicating higher quality assemblies from experiments E01, E07, E09, E10, and many samples from E02 and E06. The poorest quality assemblies are the E03 liver hypnozoite samples, with less than three BUSCOs found complete/fragmented for both samples with both BUSCO datasets. Other samples in the lower left quadrant of the graph include many asexual samples which did not assemble well from E02, E04, and E06. These may explain the low recovery of *pir*s from where we expect them to be found

Reference	Life cycle stages	Strains used	Experiment code
(Zhu et al., 2016)	IDC (different lengths of time in *ex vivo* culture)	NW Thai patient isolates (Mae Sot)	E01
(Kim et al., 2017)	Mixed blood stages (and one sporozoite)	Ratanakiri, Cambodia	E02 (Same authors as E06)
(Gural et al., 2018)	Mixed liver stages and 'hypnozoite-enriched' (*in vitro* human organelle system)	Thai patient isolates	E03
(Cheng et al., 2019)	Mixed blood stages	5 N Thai and 5 S Thai patient isolates	E04
(Muller et al., 2019)	Sporozoites	Thai patient isolates (Tak/Ubon)	E05
(Kim et al., 2019)	Mixed blood stages (with differential counts)	26 Ratanakiri, Cambodia patient isolates	E06 (Same authors as E02)
(Gunalan et al., 2019)	Mixed blood stages from two species of monkey	*P. vivax* Sal I (El Salvador origin (Collins et al., 1972))	E07
(Boonkaew et al., 2020)	Ookinete (18h) and Oocyst (7days)	Thai patient isolates (Ubon and Yala)	E08
(Siegel et al., 2020)	Schizonts	Cambodia patient isolates	E09
(Rangel et al., 2020)	IDC (4, 20, 36, 44 and 72 hours post-thaw)	Acre, Brazil	E10

Table 5.3 – Summary of the publications from which *P. vivax* RNAseq data was taken and used for *de novo* assembly.

The study reference, life cycle stages/conditions extracted, geographic source of the isolates and experimental code used in this thesis are all included.

in high quantities (smallest *pir* recovery from asexual blood stage samples was zero in one sample each from E02, E04 and E06). Mosquito stages show low *pir* recovery but a mixed pattern of BUSCO recovery. The E08 samples (including ookinetes and oocysts) show reasonable Apicomplexan BUSCO levels but poor *Plasmodium* BUSCO levels, outliers to the general trend of these two BUSCO scores being correlated. Most of the sporozoite samples of E05 are reasonably well assembled, however they still have low numbers of *pir*s (0-5 *pir* mRNAs), with one exceptional sample of 84 *pir* transcripts. The relationship of *pir* transcript discovery to total transcripts assembled is shown in Figure 5.4. In the bottom left of this figure are assemblies which have low total numbers of transcripts recovered, including some as few as ~2000. These include the E03 hypnozoite-enriched liver samples which also showed negligible BUSCO recovery.

It is impossible to conclude from the low-read RNAseq datasets whether they are genuinely lacking in *pir* transcription since the total amount of the transcriptome assembled is tiny. This may also include the E03 liver schizont samples, although these transcriptomes are several thousand transcripts larger than the hypnozoites and contained a higher proportion of BUSCOs. We have better evidence of a dearth of *pir* transcription in the datasets which produced higher numbers of total transcripts and BUSCOs, but negligible *de novo pir* transcripts, such as the E08 mosquito stages and most of the E05 sporozoite samples. This matches observations from *P. berghei* that *pir* transcription is mostly absent in the analysed mosquito stages of development. There is, however, the one exceptional sample from E05 that gave rise to tens of *pir* transcripts, hinting that this paradigm is not absolute. IDC *de novo* transcriptomes can be found all over the figure, representing results too small to make conclusions from (particularly from E06), results with high numbers of transcripts but low *pir* recovery (many from E04), results with average numbers of transcripts but high *pir* recovery (E01, E02 and E06 samples, and the schizonts of E09), and finally the E07 assembly which has high overall transcriptome and *pir* transcriptome size. The fact that the greatest *pir* recovery is obtained from the asexual blood stages mirrors both *P. berghei* and *P. c. chabaudi*. There is high variation of the number of *pir* transcripts found with reference to the total *de novo* transcriptome.

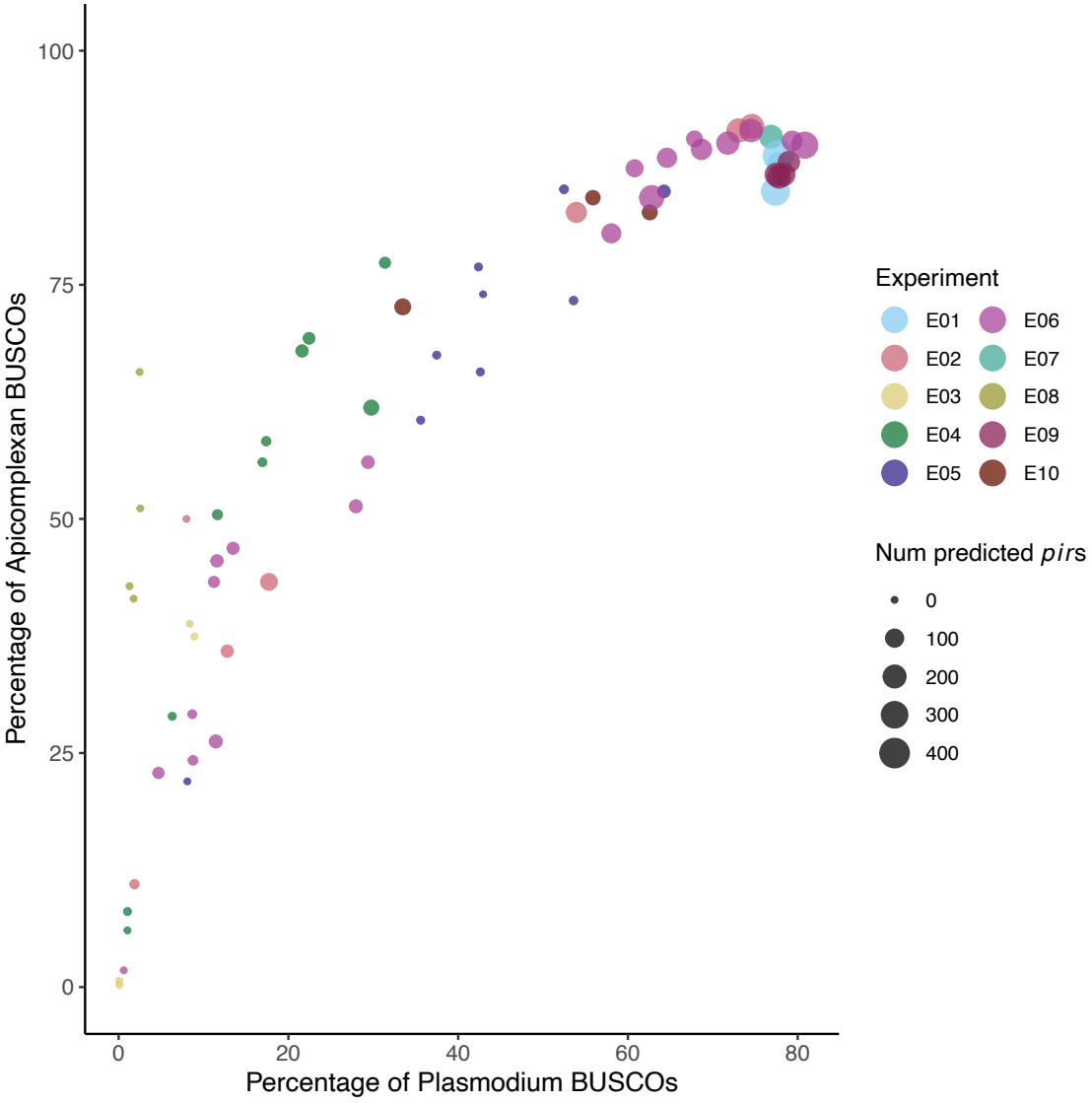


Figure 5.3 – Percentage of BUSCOs present in the assemblies.
BUSCOs are conserved genes that are expected to be found in assemblies from given organisms, in this case the Apicomplexan (y-axis) and Plasmodium (x-axis) BUSCO datasets were used. BUSCOs found includes 'Complete', 'Duplicate', and 'Fragmented'. Colour corresponds to the experiment of origin of the assembly, and size indicates the number of predicted *pir* transcripts (with ORFs over 200aa in length). Note that the x axis is shorter than the y axis.

This could be a consequence of the stage composition when samples were taken from the circulation, as results in this thesis and elsewhere have demonstrated that the delineated classes of cells of the IDC have different amounts of *pir* transcription in the rodent malarias. However, this does not explain why several of the asexual blood

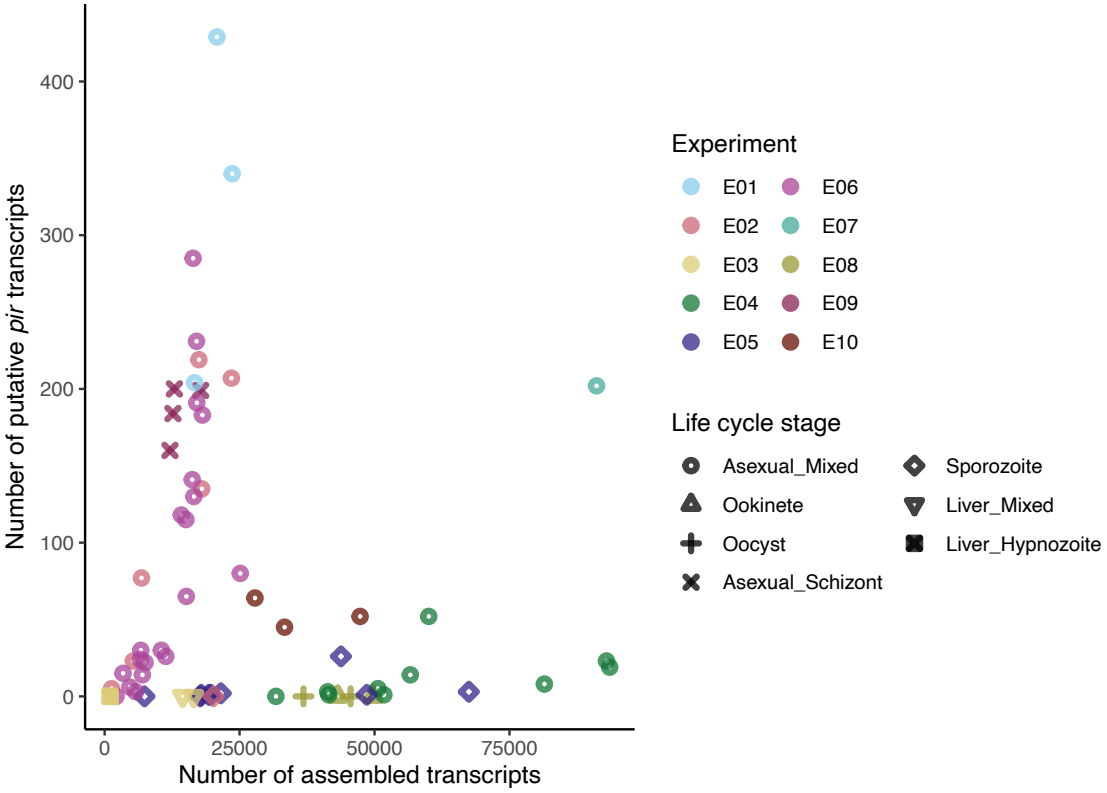


Figure 5.4 – The number of *de novo pir* transcripts with predicted *pir* ORFs at least 200aa long against the total number of transcripts for different *de novo* assemblies of *P. vivax* RNAseq.
The experiment is shown by the colour, and the stage of the life cycle the RNAseq samples are derived from is shown by the shape.

assemblies have some of the highest numbers of total transcripts but well under twenty *pir* transcripts. The lone assembly in the top right of the graph, which possesses one of the largest *de novo* transcriptomes with ~200 putative *pir* transcripts, is from E08 which used the *P. vivax* Sal1 strain. This is a reference strain of the parasite, which is probably more clonal than wild isolates after years of being passaged through monkeys (Collins et al., 2009). Having a less heterogeneous pool of sequences may make the assembler more efficient with the construction of gene transcripts.

To further validate these *de novo* transcripts, I aligned the original RNAseq data to the assemblies. This permitted me to ascertain which of the transcripts with *pir* ORFs can be mapped directly onto the original RNAseq reads. With such a variation in the size of the *de novo* transcriptomes, I was concerned that this would impact the TPM values of transcripts after mapping and quantification (since TPM for a given sample always sums

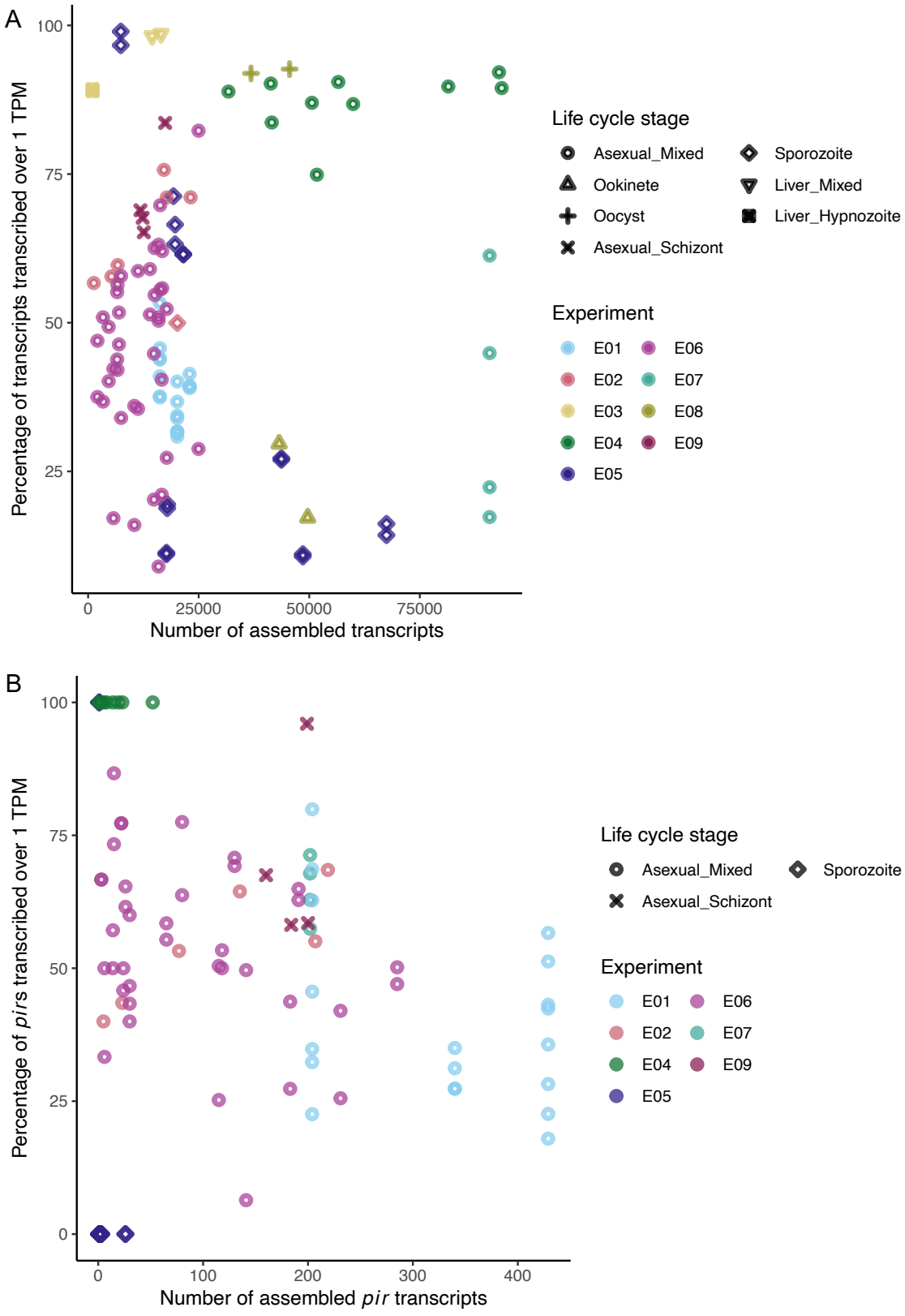


Figure 5.5 – Percentage of the *de novo* transcriptome transcribed over 1 TPM against the total size of the transcriptome.

Shown for all genes (A) and subsetting by the *pir*s (B). Different experiments are shown by colours and life cycle stages by shape. Each point represents one biological replicate of RNAseq mapped against its cognate assembly. Some replicates were combined to construct assemblies (see Materials and Methods) so these will have the same x-axis position. Note that E10 was not used for transcription quantification due to an undiagnosed software error.

to 1 million, with more mapping reference sequences individual TPM values are likely to be diluted, and with a small reference they may be inflated). However, plotting the percentage of the transcriptome which is transcribed above 1 TPM in each of the samples shows that there is no relationship between the size of the *de novo* transcriptome and the proportion of said transcriptome found to be actively transcribed (see Figure 5.5A).

This is the same for the *pir* portion of the transcriptome, perhaps with a slight negative correlation due to more lowly-transcribed *pir*s being assembled (see Figure 5.5B). The assemblies with high transcript numbers but poor mapping of RNAseq reads may be indicative of sequencing libraries/transcriptomes with low complexity, which may be the case for the sporozoite and ookinete samples in Figure 5.5A.

I applied a threshold of each *pir* transcript being transcribed above 1 TPM in at least one sample to the original set. This selection for evidence of transcription trims the original 4485 total number of putative *pir* transcripts to 2833 (possessing 2889 ORFs). Notably this removes almost all the handful of potential *pir* transcripts from sporozoite samples, with the exception of one from E05 (which was not the sample with a high number of *pir*s noted above).

Overall, I found 2833 *de novo pir* transcripts (2889 ORFs) in the assemblies with transcriptional evidence. These were massively enriched in the asexual blood stage samples, and virtually absent from mosquito stages. The liver hypnozoite samples, in particular, had so few assembled transcripts that this stage is likely under-sampled and hence I cannot conclude about *pir* abundance at this point of development.

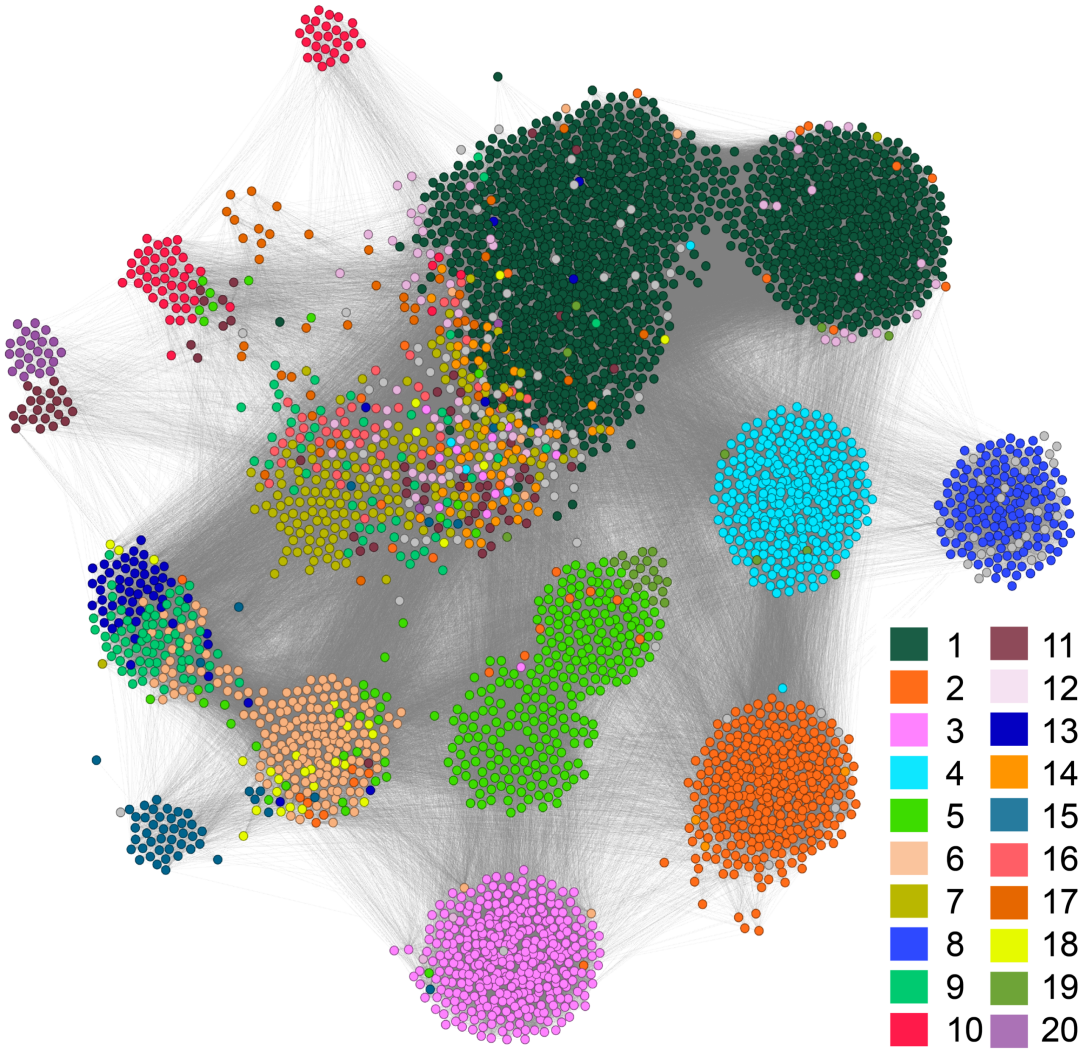


Figure 5.6 - Visualisation of the 30% identity and 25aa length similarity PIR network, showing the top twenty clusters, using an MCL inflation value of 1.4.
The rest of the 33 total clusters are coloured grey. Only the main component is shown, a second three-member component is not present. Layout was produced using OpenOrd in Gephi. Each node corresponds to a PIR ORF. Singletons are not included. Colours correspond to the top 20 MCL cluster assignments, with all other clusters shown as 'grey'.

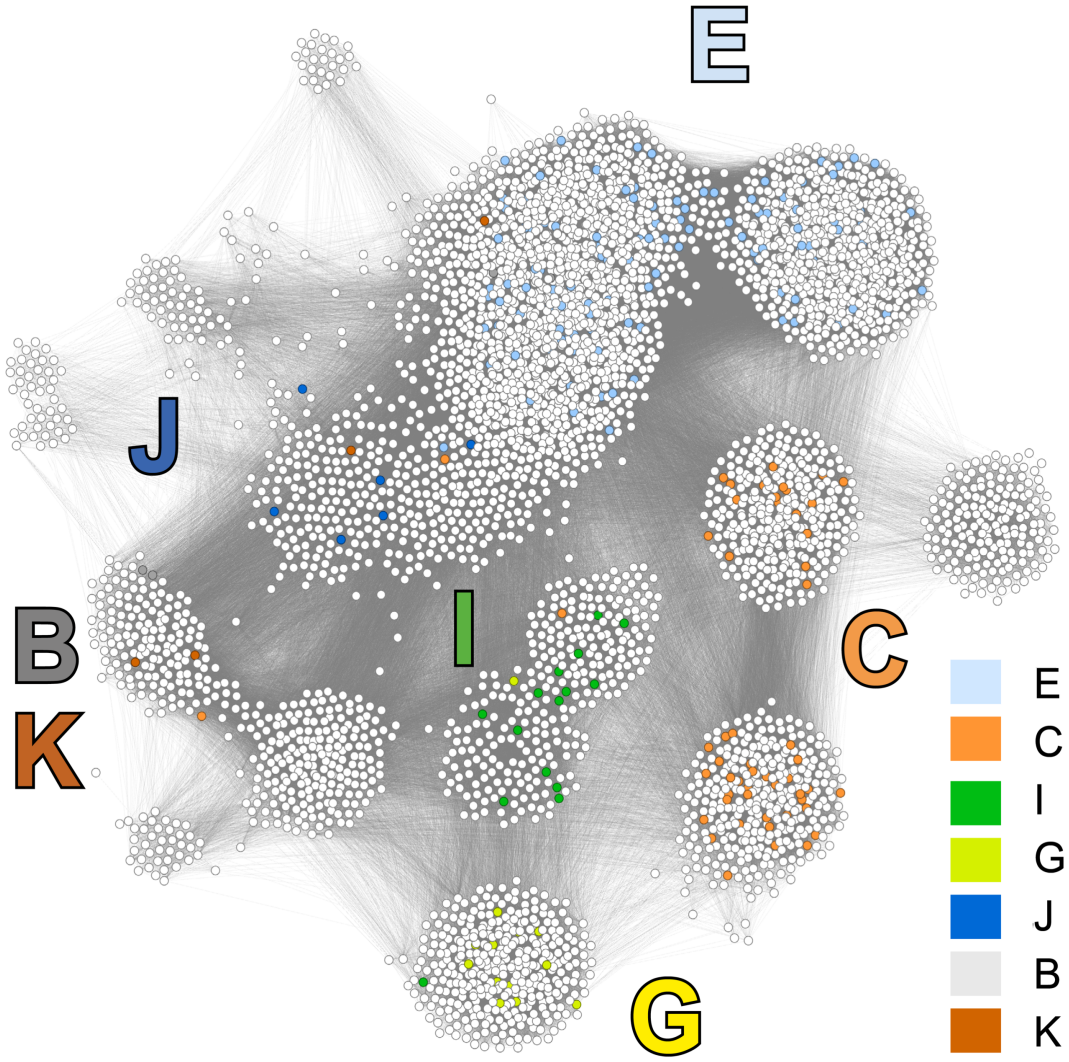


Figure 5.7 - Visualisation of the 30% identity and 25aa length similarity PIR network, showing the assigned Sal1 sub-family sequences.
All other sequence nodes are coloured white. Only the main component is shown, a second three-member component is not present. Layout was produced using OpenOrd in Gephi. Each node corresponds to a PIR ORF. Singletons are not included. The major sub-families of different visible clusters in the network are highlighted by the letters overlaying the figure.

5.3.3 *De novo pir* transcripts both cluster within previously described sub-families and describe novel groups

Using the 2889 *pir* ORFs of at least 200aa in length, and transcribed by at least 1 TPM in one sample, I produced a similarity network of BLAST alignments. Also included in this network was the 1216 *P. vivax* P01 PIR protein sequences and 247 *P. vivax* Sal1 PIR protein sequences identified from a search of PlasmoDB (Aurrecoechea et al., 2009), as well as 116 *de novo* ORFs identified by the initial Zhu et al., 2016, study (Zhu16). These 'known' sequences permit comparison of my results with those of previous investigations. Importantly the Sal1 sequences were used to define the sub-families of *P. vivax pir* genes (sub-families B, C, E, G, I, J and K), so we can investigate the assignment of sub-families to the *de novo pir*s (del Portillo et al., 2001; Lopez et al., 2013). By performing BLAST on this collection of protein sequences (the 2889 *de novo* PIR ORFs and the 1579 published PIRs, giving a total of 4468 PIRs in the network) against itself I created a similarity network which can be used for clustering. I tested several different given thresholds of BLAST coverage and percentage identity, as well as 'inflation' values of the MCL clustering algorithm, to see which produced the best clusters of the *pir* genes. The optimal network/clustering would include as many of these 4468 total PIRs as possible, have a number of clusters which enables meaningful analysis, and also minimise mixing of the previously defined sub-families. Comparing the different parameters found that a set of thresholds defining edges by a minimum 25aa long BLAST alignments with at least 30% identity produced a network which included 4423 (99.0%) of the input sequences. This network was composed of almost entirely one component, however MCL clustering (with an inflation value of 1.4) assigned 59 clusters (32 with more than two members) within this structure. The only separate component from the main conglomeration was made up of three Zhu16 sequences.

To obtain visual verification of the integrity of the clustering assignment, I constructed the network within Gephi visualisation software (Bastian et al., 2009). Even though this highly interconnected network forms a 'hairball', the network layout algorithm employed

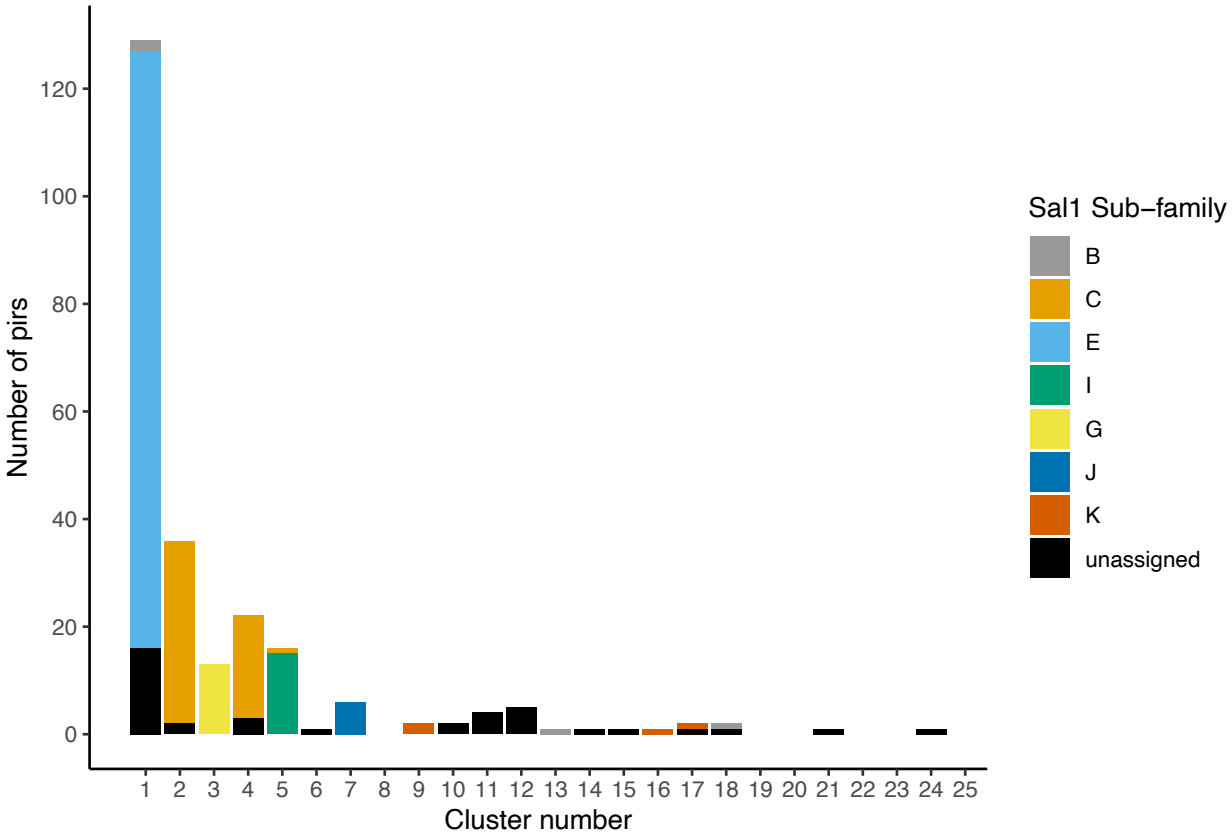


Figure 5.8 – The distribution of the Sal1 sequences across the largest twenty-five clusters of the 30% identity and 25aa length similarity PIR network, clustered using an MCL inflation value of 1.4.

Colours show the sub-family assignment of the PIR sequences, including those which were 'unassigned' in black. Those clusters without any Sal1 sequence are included without any bar in the chart.

Cluster	Num. PIRs	Sal1 sub-family sequences	Sub-family assignment
1	1826	B (n = 2); E (n = 111)	E
2	362	C (n = 34)	C_1
3	358	G (n = 13)	G
4	314	C (n = 19)	C_2
5	295	C (n = 1); I (n = 15)	I
6	185	N/A	unassigned_1
7	166	J (n = 6)	J
8	146	N/A	unassigned_2
9	118	K (n = 2)	K_1
10	73	N/A	unassigned_3
11	72	N/A	unassigned_4
12	65	N/A	unassigned_5
13	59	B (n = 1)	B_1
14	56	N/A	unassigned_6
15	53	N/A	unassigned_7
16	42	K (n = 1)	K_2
17	42	K (n = 1)	K_3
18	29	B (n = 1)	B_2
19	28	N/A	unassigned_8
20	26	N/A	unassigned_9
No BLAST alignments	*45*	*N/A*	*N/A*

Table 5.4 – The largest twenty clusters from the 30% identity and 25aa length similarity PIR network, clustered using an MCL inflation value of 1.4. Includes the total size of the cluster, the number of Sal1 sub-family sequences in each, and the sub-family assigned to each sequence in the cluster based on this. The last row includes information on the 45 PIR sequences not included in the network (noCluster). Colours correspond to those used to distinguish these sub-family *pir*s in the transcriptional graphs (Figure 5.10, Figure 5.11, and Figure 5.12).

was able to find structure which strongly corresponded to the MCL assignments (see Figure 5.6). Overlaying the Sal1 sub-family sequences onto this network showed that they too segregated neatly into different clusters (see Figure 5.7). Perhaps the biggest disagreement between the Gephi network layout and the MCL clustering is that the largest cluster (E sub-family) may be more accurately described as two separate but highly connected groups. Cluster 18 (yellow) contained the sporozoite PIR, and it is

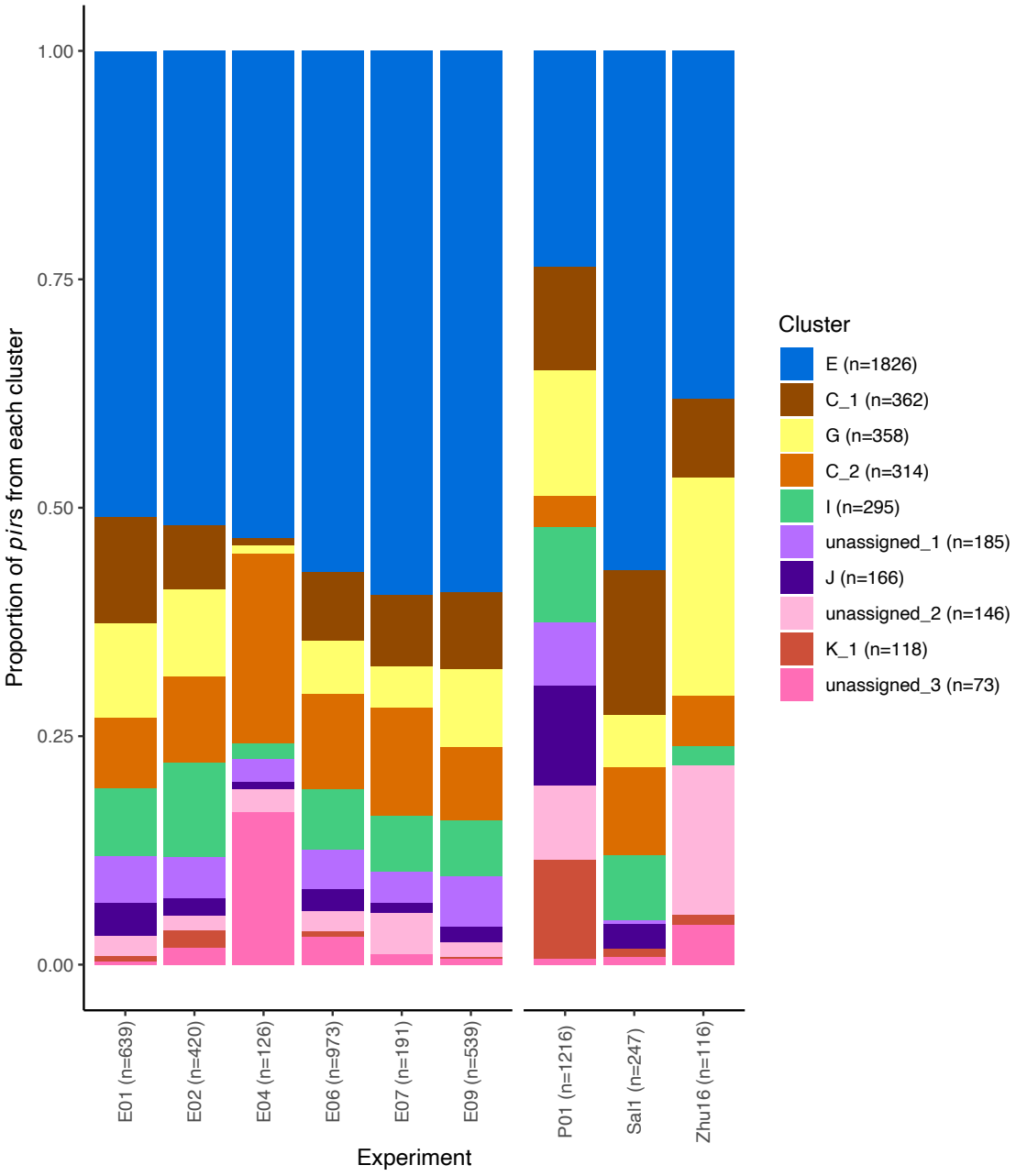


Figure 5.9 – The composition of the largest ten clusters among the experimental and known *de novo* PIRs sequences from the 30% identity and 25aa length similarity PIR network, clustered using an MCL inflation value of 1.4.
Cluster labels show the total number of different sequences in each one, and the Experiment labels show the total number of *de novo* sequences from each. The right hand side of the graph shows the 'known' PIR sequences from the P01 and Sal1 reference genomes and the original Zhu16 *de novo* transcripts.

noteworthy that this cluster does not show any structure in the Gephi visualisation. Clusters around the same size, such as 30 members or fewer, tend to show little similarity to the Gephi visualisation and should be interpreted with caution.

Previous *pir* sub-family analysis has identified distinct clades that may be classified into longer or shorter groups based on their average peptide lengths. Otto et al., 2014, showed that the phylogeny of rodent malaria split into the Long-forms and the Short-forms, with the former group's subdivisions ranging from an average length of ~1000-2400aa and the latter group's subdivisions ranging from an average length of ~800-950aa. To see whether this pattern can be observed among the *P. vivax pir*s, I have analysed the median PIR ORF length of the *P. vivax* clusters - with the Zhu16 *de novo* ORFs removed as they tended to be particularly short. Cluster 8 contains the longest average sequences, with a median of 482aa, which is shorter than the shortest average of rodent malaria Short-forms. To determine that these *P. vivax de novo* transcripts are full-length would require identification of C- and N-termini of the predicted peptides, so it is possible that these medians are skewed down by fragmentary sequence. Even with this caveat, I observe that longest *P. vivax* PIRs overlap only with the Short-form PIRs, and no sequence extends into the lengths of the Long-forms, so they do appear to be consistently shorter than rodent parasite PIRs. Furthermore, there is no obvious split between longer and shorter groups, suggesting that the Long-form and Short-form structure has no equivalent in the *P. vivax*-lineage *pir* family, or even that all *P. vivax pir*s are an expansion of an ancestral Short-form family.

Among the sequences within these clusters there was almost no mixing of different Sal1 sub-families (see Figure 5.8). The E sub-family, the largest, was the only sub-family in the biggest cluster of 1826 sequences, except for two B sequences. This was also the only cluster to contain E sequences. The two other B sequences, of the four in total, were found in separate clusters with no other sub-family sequence, suggesting that the B designation is not a particularly accurate sub-family on its own, and is split between the E family as well as up to two distinct other groups. Aside from the largest cluster, the only other sub-family mixing was a singular C sequence in the cluster dominated by the I sub-family. It is surprising that there is not more mixing of the C or the E sub-family given that they are the largest sub-families. The rest of the C sequences were found in the second and fourth largest clusters, with the third largest cluster containing only G sub-family members. This suggests that the C sub-family can be split into two individual

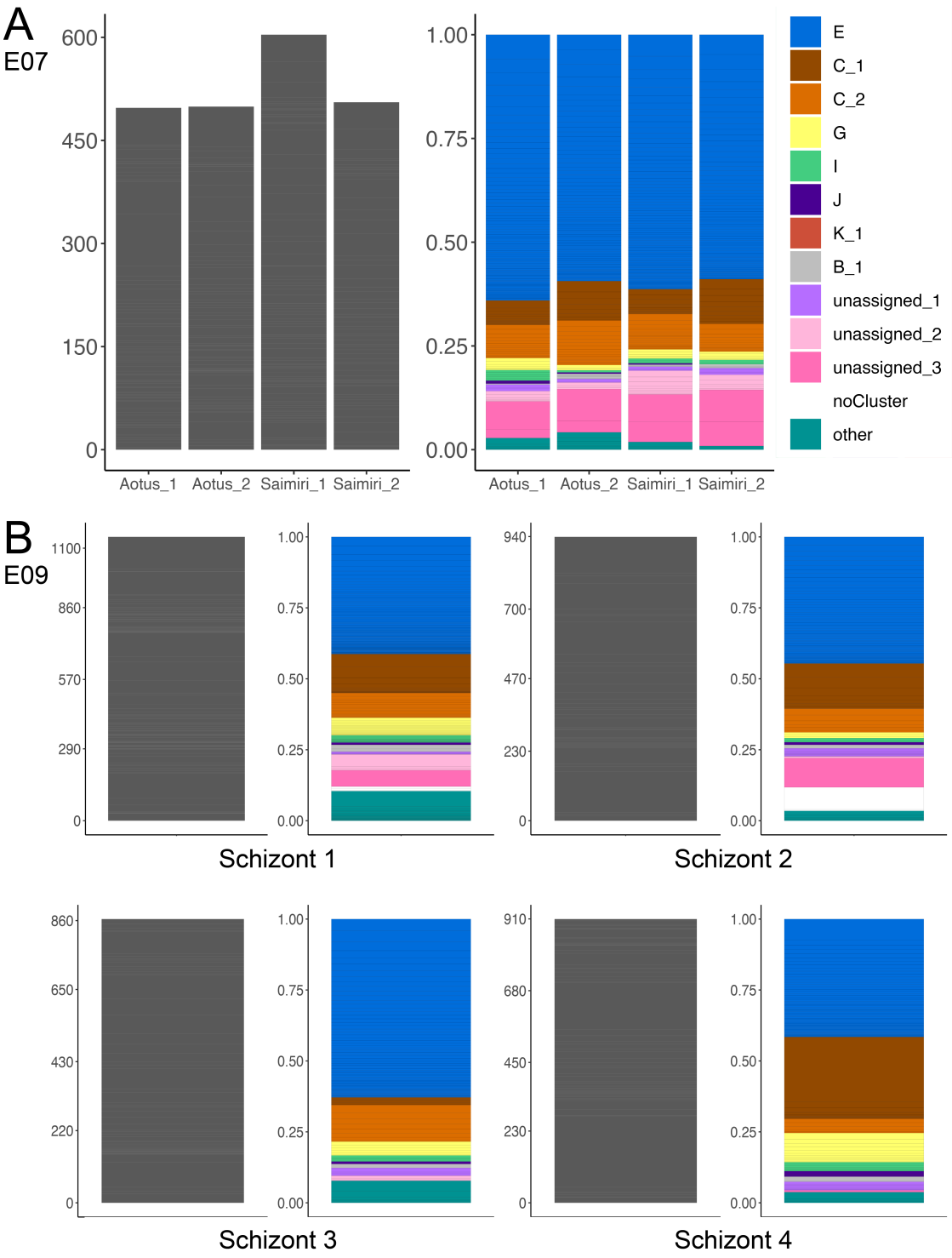


Figure 5.10 – Transcription of *de novo* assembled *pir* genes from RNAseq alignment to the assemblies for E07 and E09.

Graphs on the left show the absolute TPM of all *pir*s and on the right show proportions of the total *pir* TPM contributed by individual genes and coloured by their sub-family. Each pairing of absolute/proportional TPM graphs denotes a given assembly, which may be constructed from, and mapped against, multiple samples

(as in A) or just one (as in the B graphs). A shows the results of E07 (Sal1 infection of *Aotus/Saimiri* monkeys), as well as the key for sub-family colours, and B shows the results of E09 (patient isolate schizonts). The sub-families which were coloured were selected by size and interest, all other sub-family/cluster *pir*s present are grouped as 'Other' (turquoise).

groups, C_1 and C_2. The C_2 sub-family includes the ancestral *pir* genes from P01 and Sal1, as well as >50% identity BLAST matches of *de novo pir*s from the E02, E04, E06 and E07 assemblies – presumably these isolates' ancestral *pir* sequences constructed from the RNAseq reads. It is curious that only a few experimental assemblies contained high quality ancestral *pir* sequences, however perhaps other assembly tools would pull this particular gene from the RNAseq reads for the other isolates, as seen with *P. c. chabaudi* AS data. The major sub-families that are found in each cluster can then be used to assign a sub-family to other sequences within the same groups, as identified in
Table 5.4. In clusters smaller than cluster 9 (K_1) the sub-family associations are based on single sequences, which makes them spurious. These network thresholds excluded 45 total PIR sequences which lacked the requisite similarity to one another. Twenty of these sequences were not found to be similar to any other sequence by the initial BLAST alignment (maximum 50 matches at least with an E value of $1x10^{-3}$) and are likely divergent *pir* sequences, fragments or mis-annotations.

Observing the spread of the cluster assignments across the *de novo* PIRs from the experimental RNAseq data I note that they are generally mixed throughout. For the largest ten clusters there is a *de novo* PIR in almost every cluster from each experiment (see Figure 5.9). This is reassuring for the clustering and classification process, as it shows that the variation within the data is not only due to the experimental source. This also matches previous studies which have shown extensive overlap between PIR sequence clusters of different isolates (Minassian et al., 2021). The larger sources of PIRs tend to have representation in all the top twenty largest clusters, including from E01, E02, E06 and the P01 reference genome. A particularly large proportion of P01 PIRs are members of cluster 9 (K_1), while P01 is depleted from clusters 1 (E), 4 (C_2) and cluster 10 (Unassigned_3). The relatively small number of E sub-family members in P01 makes this genome assembly stand out, as it is by far the largest sub-family overall. Unassigned_3 has the highest proportion of *de novo* PIRs among the largest twenty

clusters, with no assigned Sal1 sub-family members, a striking enrichment of E04 *de novo* PIRs (which is absent from many other clusters (see Figure 5.9)), and conspicuously few E01 sequences.

The similarity of the transcribed *de novo* PIR sequences and those from reference genomes demonstrates strong overlap. Clusters associate with most of the Sal1-reference sub-families, although some clusters require new classification, most notably Unassigned_3 which is under-represented by the published genomes included in this thesis.

5.3.4 Transcription of *de novo pir*s is both spread over the major sub-families and diverse between parasite isolates

Finally, I used the Salmon RNAseq alignments of the original samples to observe the patterns of *pir* sub-family transcription within and between isolates/conditions. Many of the assemblies with low BUSCO scores showed much lower *pir* diversity across the transcriptional profile compared to those with higher numbers of *pir*s and better BUSCO scores. For this reason, I focussed on showing the results from assemblies which had BUSCO recoveries >50% for both the Apicomplaxan and *Plasmodium* datasets. This excludes all the E04 samples, which gave rise to comparatively poor assemblies despite recovering *pir* genes. This also removes the E05 sporozoite sample which transcribed a singular *pir* gene at low levels (<10 TPM).

As the largest sub-family, it is not surprising that the E *pir*s constitute the highest transcribed group across most samples, however the exact proportion can vary substantially. In Figure 5.10 I contrast the transcription of the *pir*s in both E07 (A – infections of splenectomised simian hosts with the Sal1 reference isolate) and E09 (B - schizont samples from different patient isolates). In Figure 5.10A we see that the different hosts do not majorly impact *pir* gene transcription, both in terms of absolute transcription levels and sub-family proportion. In this experiment it is reasonable to say that the source isolate is determining the *pir* transcriptional patterns; over 50% E *pir*s (blue), and high C_1 (brown) and C_2 (orange) transcription. Figure 5.10B shows separate isolates/assemblies, with more substantial variations in the sub-family

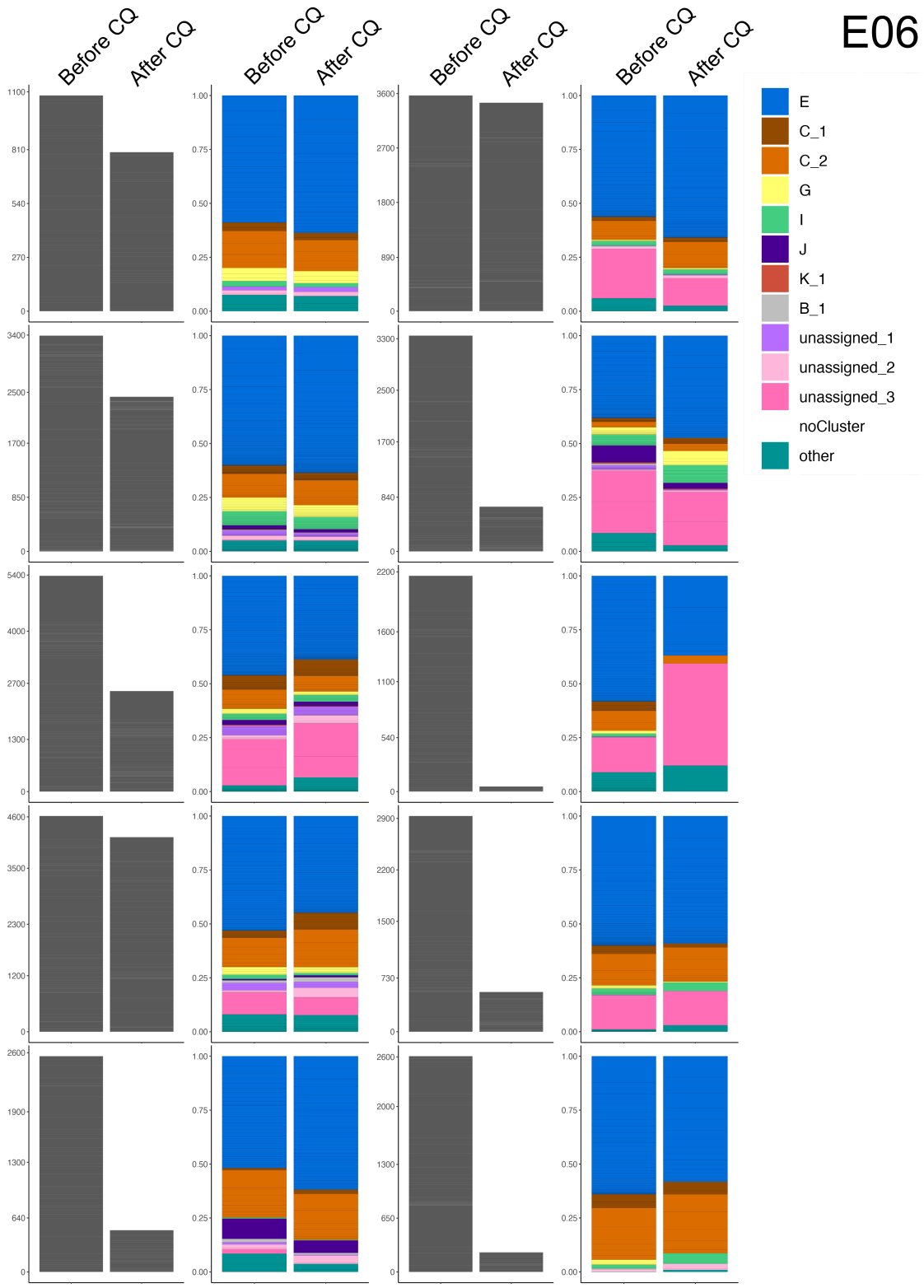


Figure 5.11 – Transcription of *de novo* assembled *pir* genes from RNAseq alignment to the assemblies for E06.

Graphs on the left of each pair show the absolute TPM of all *pir*s and on the right show proportions of the total *pir* TPM contributed by individual genes and coloured by their sub-family. Each pairing of absolute/proportional TPM graphs denotes a

given assembly constructed and mapped against a pair of samples (before and after chloroquine (CQ) treatment. The sub-families which were coloured were selected by size and interest, all other sub-family/cluster *pir*s present are grouped as 'Other' (grey).

proportions between them. In this case only one sample (Schizont 3) has over 50% E *pir* transcription, at the expense of much depleted C_1 transcription. Schizont 4 shows C_1 transcription rivalling that of the E sub-family, and Schizont 2 has a major component of the transcriptional profile from a single *pir* that is dissimilar to the others in this analysis (marked as noCluster - white). Across both Figure 5.10A and B, a relatively small cluster that frequently appears to be highly transcribed is Unassigned_3 (hot pink). In E07 two *pir*s from this sub-family are transcribed at over 10% of the total *pir* transcription, while in E09 two separate schizont samples (1 and 2) have a prominent band of Unassigned_3 transcription from individual *pir*s. Note that this corresponds to cluster 10 in Figure 5.9, a cluster with particularly low representation from the reference genomes of P01 and Sal1. Together these figures demonstrate a common pattern of lower intra-isolate variation compared to inter-isolate. Although different isolates can vary substantially from one another, it is still worth noting that many patterns remain the same, such as most transcription coming from the larger sub-families (particularly E and C_2), and most of the major sub-families being transcribed at least a little bit.

The above patterns are replicated by the E06 data, which looked at patient isolates before and after the administration of chloroquine (CQ). The different isolates show strong variation to one another, while variation of sub-family proportions of the same infection between therapeutic conditions is less stark (see Figure 5.11). Some examples of this are the strong C_1, G (yellow), I (green), or J (dark blue) sub-family signals present in some samples but being absent in others. The most conserved behaviours appear to be strong E sub-family signatures and at least some C_2 transcription. Once again Unassigned_3 is represented by a small number of *pir*s being transcribed at high proportions in most of the patient isolates in both conditions. The absolute values do show a major change of *pir* TPM values after the *P. vivax* infections are treated with chloroquine, and this response itself varies greatly between isolates. The original study also observed this differential expression across their samples when aligning to a reference genome, and since overall transcription (by PCA) does not show major

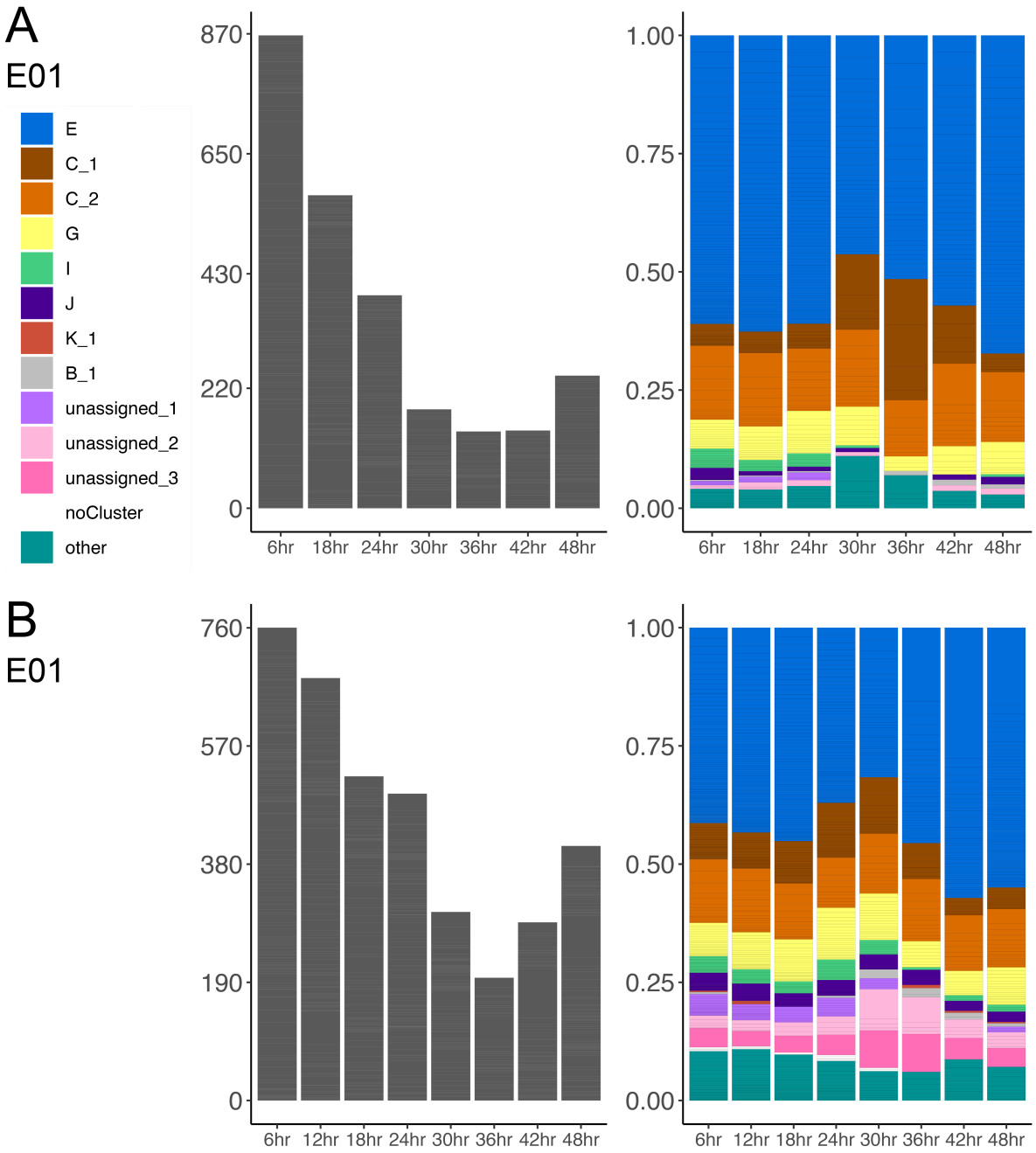


Figure 5.12 – Transcription of *de novo* assembled *pir* genes from RNAseq alignment to the assemblies for E01 across the IDC, including patient isolate 1 (A) and patient isolate 2 (B).

Graphs on the left show the absolute TPM of all *pir*s and on the right show proportions of the total *pir* TPM contributed by individual genes and coloured by their sub-family. Each pairing of absolute/proportional TPM graphs denotes a given assembly constructed from and mapped against the samples taken across the IDC for each isolate. The sub-families which were coloured were selected by size and interest, all other sub-family/cluster *pir*s present are grouped as 'Other' (turquoise).

Original publications describe 6hr as 'early ring', 18hr as 'late ring', 24hr as 'early trophozoites', 30hr as 'late trophozoites', 42hr as 'mid schizont', and 46hr as 'late schizont' (Bozdech et al., 2008; Zhu et al., 2016).

variation caused by the chloroquine the *pir*s were one of only few gene sets to majorly respond to drug treatment (Kim et al., 2019). Once again, most transcriptional profiles involve many *pir* sub-families dominated by the Es with strong C_2 contribution.

Having established that asexual blood stage transcription of *pir*s is high, and that sub-family distributions of transcription are influenced majorly by isolate, I looked at the E01 results from across the IDC (see Figure 5.12). These results also show transcriptome changes which can be compared to that of *P. c. chabaudi* AS from Chapter 4. Both Figure 5.12A (SMRU1 isolate) and B (SMRU2 isolate) demonstrate a pattern of *pir* transcription peaking in early rings (earliest time point is 6hr), troughing around early schizonts (36hr), before showing signs of upregulation by the final timepoint, late schizonts, of the complete cycle. Given that the first timepoint used is 6hr into the *ex vivo* incubation, it is possible that an earlier timepoint would correspond better the end of the cycle shown. However, by the final timepoint the cells will be struggling to grow, as *P. vivax* cannot be cultured *ex vivo* for long, so low *pir* transcription at the end of the cycle compared to the start may be due to cellular stress. Comparison with *P. c. chabaudi* AS shows that, although both species demonstrate unimodal distribution of *pir* transcription across the IDC, the timings of peaks and troughs are different. For *P. vivax, pir* transcription appears to peak when it is lowest in *P. c. chabaudi* AS (early ring stages) and troughs when it is highest in *P. c. chabaudi* AS (late trophozoites/schizonts). These results correspond with those of the original microarray and RNAseq investigations based on these samples (Bozdech et al., 2008; Zhu et al., 2016). There are compositional changes of the sub-families across the IDC as well as the absolute TPM changes. Early rings to late trophozoites show some transcription of the I sub-family, while this is absent by the time schizonts appear. Smaller groups of *pir*s expand in their proportional transcription when *pir*s overall are downregulating, suggesting that changes in the major families of *pir*s, such as the Es, are the major drivers of IDC transcriptional changes. In Isolate 1 the transcription of the E-family by 30h (just before *pir* transcription is at its minimum) is replaced mostly by the C_1s and C_2s, while in Isolate 2 it is replaced by Unassigned groups 2 and 3. Once again, the two isolates have variation in the *pir* sub-families transcribed, with the second isolate showing more

diverse transcription even though the two isolate assemblies have similar *Plasmodium* BUSCO scores (Figure 5.3).

When *pir*s are transcribed highly in *P. vivax* samples, they tend to include the major sub-families even though the proportions of these sub-families vary. When the same isolate is used for infections of different hosts or across different conditions and time-points of asexual blood stage development, the *pir* sub-family repertoire stays mostly the same, with only subtler changes observed. It is between isolates that most variation can be observed, with large changes in the proportions of different sub-families even if the larger families remain expressed. Total *pir* transcription shows a unimodal distribution across the IDC of *P. vivax*, however by peaking in early rings/late schizonts it demonstrates the opposite pattern to that observed in *P. c. chabaudi* AS.

5.4 Conclusions

Although they may be relevant for human infections, transcription of the *P. vivax pir* genes can frustrate study due to their diversity between isolates and reference genomes. To overcome this, I used *de novo* transcriptome assembly to construct the *pir*s from the RNAseq data and then apply these as a mapping reference. After testing on *P. c. chabaudi* AS data, which we used as a 'gold standard' laboratory strain with well-annotated *pir*s which should vary minimally compared to the genome reference, I proceeded with SPAdes single-cell mode for the *P. vivax* assemblies. The results of these assemblies and mapping back to the original RNAseq reads demonstrated that hundreds of *pir*s can be constructed by this method from individual samples. Mosquito samples displayed negligible *pir* recovery and asexual blood stages had high numbers of *de novo pir*; results which match the rodent malaria models. Analysis of the similarity network of the ORFs of these *de novo* sequences and reference genome sequences showed high intermingling of sequence from different experimental sources, and close matching of clusters to existing sub-family definitions with a few additional groups not covered by the current definitions (Lopez et al., 2013). Transcriptional profiles show that many of the predicted sub-families are transcribed in infections, and proportions of sub-family transcription from the same isolate tend to remain minimally unaffected by different hosts and drug treatment. There is stronger variation of these proportions between different isolates, suggesting that *pir*-transcription is isolate-specific and not random upon infection. Analysis of transcription across the IDC suggests that

expression in *P. vivax* is high, as seen in *P. c. chabaudi* AS, but *pir* expression peaks and troughs appear at different times.

Despite being one of the most common tools (and the default for many bioinformaticians) and assembling the largest number of transcripts in total, Trinity performed poorly with recovering *pir* genes. This suggests that the previous *de novo* assembly studies in *P. vivax*, which used Trinity, may have underestimated the *pir* repertoire (Kim et al., 2017; Zhu et al., 2016). Trinity's algorithm ranks all the k-mers in a dataset before assembling based on the upper proportions, so this may have filtered out the low frequency k-mers of *pir* gene transcripts. The SPAdes assembly worked considerably better, however this threw another surprise when the single-cell algorithm appeared to return better results than the dedicated *de novo* RNAseq algorithm. The superior performance of sc-SPAdes has been observed before, particularly for yeast transcriptome assembly which may suggest that it would be preferable for other unicellular eukaryotic organisms (Hölzer & Marz, 2019).

No individual assembler works better than all others in every metric for each data source (Amin et al., 2014; Chopra et al., 2014; Francis et al., 2013; Hölzer & Marz, 2019; Madritsch et al., 2021; Rana et al., 2016; Yahav & Privman, 2019). Hence, using multiple *de novo* transcriptome assemblers and combining them into a high-quality master assembly would be a preferable method to utilise the expertise of each tool and obtain comprehensive assembly. Such an approach could incorporate assemblers not tested in this thesis but which produce high quality *de novo* transcriptomes, such as Bridger (Chang et al., 2015), Trans-ABySS (G. Robertson et al., 2010), and SOAPdenovo-trans (Xie et al., 2014). Some of these approaches also accept different k-mer length parameters for their algorithms, which should be applied across a range of values in order to obtain expansive transcript construction. Combining these assemblies together requires a redundancy-removal step, such as that performed with CD-HIT in this thesis. However, CD-HIT chooses the longest transcript as the 'representative' sequence of each cluster of redundant transcripts, and this is not necessarily the most correct one (D. G. Gilbert, 2019b). Tools such as EvidentialGene provide alternative methods of removing redundancy and extracting the highest quality representative sequences (D. Gilbert, 2016a, 2016b; D. G. Gilbert, 2019a). A downside to EvidentialGene is that the programme's methods are not well described, making it a software black-box. Other steps that could be added to the *de novo* assembly workflow

include further steps of quality control, such as those provided by TransRate (Smith-Unna et al., 2016) and rnaQUAST (Bushmanova et al., 2016). All-in-all the most comprehensive assessment of a *de novo* transcriptome with existing tools requires a huge amount of scripting and learning new commands. A pipeline software, such as Nextflow used with the *P. berghei* and *P. c. chabaudi* AS analyses in this thesis, could make this substantially easier to obtain (di Tommaso et al., 2017). There are some existing Nextflow pipelines for *de novo* transcriptome assembly, such as TransPI (Rivera-Vicéns et al., 2022). I was unable to make that particular pipeline operational, so perhaps a Nextflow pipeline built to the standardised specifications of the nf-core pipeline database is the best next step (Ewels et al., 2020).

When these techniques were applied on *P. vivax* RNAseq data the assemblies included many *pir*s, the number of which varied greatly from experiment-to-experiment. Some of these differences could be explained by the small amount of data extracted, such as that for the liver hypnozoite samples of Gural et al., 2017. Their study found that the model of human liver tissues surrounded by mouse fibroblasts produced hypnozoite commitment varying from nearly 75% to a negligible percentage of schizont load. Estimations from human liver tissue in live mouse models (*in vivo,* but still a mouse environment) suggest maximum estimates of 15 hypnozoites per 100 mm^2 of liver section (Flannery et al., 2022). By taking a maximum estimate of square hepatocyte cross-sectional area (900 μm^2) (Qin & Crawford, 2018) I derive 111,111 hepatocytes per 100 mm^2. This suggests, at what is likely a high estimate, 1 hypnozoite per 7000 hepatocytes. Given that hypnozoites are quiescent stages anyway, this shows that *P. vivax* hypnozoite RNA will be massively under-represented in the samples. Indeed, the experimenters used hybrid capture sequences based on the P01 reference to enrich for *P. vivax* genic RNA. This would reduce the number of novel sequences that my assembling could recover anyway, and while the method was able to improve mRNA recovery enough for their study it was insufficient for *de novo* transcriptomics. The same group have since used single-cell RNAseq for extracting liver stages and this presents a better opportunity for analysing *pir* transcription (Mancio-Silva et al., 2022).

The mosquito stages of Boonkaew et al., 2020 give mixed results from the BUSCO analysis, since the Apicomplexan BUSCOs results were good but the *Plasmodium* BUSCOs results were poor. This suggests that the *Plasmodium* BUSCO dataset is lacking mosquito stage transcribed genes (in comparison to IDC genes), although the

number of *Plasmodium* BUSCOs is much larger than those orthologs conserved for all Apicomplexa. The sporozoite samples had generally good BUSCO scores, so from these data it is likely that most of the vector developmental stages of the parasites do not transcribe or require *pir*. Despite this, there was one sporozoite sample, from Muller et al., 2019, which gave rise to a singular *pir* transcript that was found to have an ORF over 200aa long and also had RNAseq evidence of transcription. While the assembly did not reach the 50%/50% BUSCO threshold I used for filtering transcription profiles of interest to show, the Salmon mapping evidence suggests that this is a real signal. This *pir* is a member of cluster 18 (sub-family B_2), alongside 28 other transcripts. However the network visualisation, and designation based on a single B sequence. cast doubt on the integrity of this cluster and sub-family assignment. Indeed, from this analysis the existence of the B sub-family is spurious, as half of members share the same cluster as the E sub-family. Since this sporozoite *pir* is found in one of the smaller clusters its sequence is divergent compared to most transcribed *pir*s. Transcription of individual *var* genes has been observed in late oocysts/sporozoites (Gómez-Díaz et al., 2017) and one of these was verified to have importance in hepatocyte invasion (Zanghì et al., 2018), a divergent function from the rest of the *var*s. Hence, there is a precedent for blood-stage multigene families being transcribed in sporozoites.

The highest BUSCO scores and numbers of *pir* transcripts came, unsurprisingly, from the asexual blood stages. This corresponds with the results from *P. berghei* and *P. c. chabaudi* AS presented earlier in this thesis, confirming a conserved requirement of *pir*s during the IDC. There was still some disparity across the *P. vivax* experiments in the quality of the assemblies, as many of the Kim et al., 2019 (E06) and all the Cheng et al., 2019 (E03), samples were far from the higher quality samples in terms of BUSCO scores and *pir* numbers. The Zhu et al., 2016 (E01), Rangel et al., 2020 (E10), and Gunalan et al., 2019 (E07) experiments, produced some of the highest quality assemblies, likely due to the combination of so many RNAseq samples from one isolate. This is not possible with patient isolates extracted directly from the host, so many of these samples suffer from low numbers of reads. Even among the better assemblies there was a wide range in the heterogeneity of the *pir* sub-families being actively transcribed, which complicated conclusions about which experiments/conditions had greater proportions of given sub-families. By including the low-quality assemblies which had *pir* transcription, it was clear that there was a relationship between the number of sub-families transcribed and the sample quality. A metric that could be used to analyse

whether the assemblies are capturing as much of the transcriptome as possible would be saturation curves of BUSCO scores and *pir* predicted transcripts. By sub-sampling from the RNAseq data (e.g. starting with 10% of the original reads and working up to the full library) and using each of these limited datasets for a separate *de novo* assembly, one can observe whether the BUSCO scores/*pir* numbers continue increasing as you incorporate more of the original reads, or whether they plateau. The latter result would suggest incomplete assembly due to lack of sufficient reads even in the total available library. I also hypothesise that the *pir* sub-families will suffer from drop-out in these data, in which existing groups of sequence but with low abundance can be 'missing' in the final assemblies if they are lacking the read evidence to be included. This is much like how in single-cell RNAseq sequences are often lost due to a combination of poor transcription and random factors. Analysis of these datasets may need to take drop-out into account in their statistics, such as by trying methods adapted for single-cell analysis (Jiang et al., 2022).

When the *de novo* assembled *P. vivax pir*s were combined with *pir*s from the reference genomes (Auburn et al., 2016; Carlton et al., 2008) and previously described *de novo* assembled *pir*s (Zhu et al., 2016), a network with clear sub-structure was formed, despite the use of lenient BLAST thresholds. The MCL cluster algorithm and visual inspection of the similarity network match well, although with smaller clusters (most notably cluster 14 and beyond) there is no obvious relationship between the sequences based on the network visualisation, so I focussed on the larger clusters for this thesis. Despite the lack of sequence data available when del Portillo et al., 2001 first defined sub-families of *P. vivax pir*s, they are still relevant for the expanded *pir* pool used in my analysis. The network clusters generated in this thesis associate cleanly with the Sal1 sub-family members, and only a few of the larger clusters do not fit into any existing definition. Although the network and clustering algorithm parameters included here were selected for purity of Sal1 sub-family members in clusters, it was noteworthy that the overall amount of mixing in any network was very low. Beyond being relevant for an updated set of *pir* sequences, the sub-families also showed relevance for *pir* transcription, as the majority of transcription for most samples came from *pir*s which grouped with Sal1 sub-family sequences. The largest clusters (E, C_1, G, and C_2) dominated most of the transcription profiles, as would be expected given their size. This was found across experiments, geographic locations, and conditions. Although extensive diversity exists in the *P. vivax pir*s, the defining groups are still found broadly

across isolates and are not unique to any strains in this investigation. This corroborates the results of Minassian and colleagues., 2021, which showed in a recent high-quality genome assembly from Thailand that the assembled genomic *pir*s nest with Sal1 and P01 *pir* sequence.

However, there were clusters which did not match any Sal1 sub-family. These clusters are not necessarily 'new', as unassigned Sal1 and P01 sequence was present in them, however these have yet to be defined. Cluster 10 (unassigned_3) had little Sal1 or P01 sequence and appears to be dominated by *de novo* assembled *pir*s (including those from the original Zhu16 publication). Like the assigned sub-family clusters, the largest three unassigned clusters are found across the experimental assembled *pir*s, however they are not so evenly distributed across transcriptional profiles. Although Unassigned_1 *pir*s are all transcribed in at least one sample over 1 TPM (as per the selection process) this group is under-represented in the transcriptional data, especially when compared to the smaller Unassigned clusters 2 and 3. As an example, from the IDC data of Zhu et al., 2016 (E01), the Unassigned_1 *pir*s appear in the transcriptional profiles at the start and end of the cycle, but when transcription troughs by late trophozoites it has been depleted, and for isolate 2 (Figure 5.12B) completely drowned out by Unassigned_2 and 3. Hence Unassigned_1 appears to represent a group of *pir*s with particularly low transcription levels. In contrast Unassigned_3 is frequently represented in transcriptional profiles usually only by one or two high transcribing *pir*s. Close analysis of the BLAST results may show that these *pir*s from different isolates are similar enough to be classed as orthologues and could represent conserved transcribed *pir*s, similar to the ancestral *pir* observed in previous chapters. The ancestral *pir* itself is a member of sub-family C_2, and does not form a lone, separate lineage as it does in rodent species. Hence, sequences more conserved with the ancestral *pir* have been retained in *P. vivax* compared to the rodent-infecting species, perhaps suggesting that the function performed is more essential, or that redundancy is more important, in *P. vivax*.

Overall, there is a common pattern of sub-family proportions generally being conserved in samples of the same isolate, while proportions vary between different isolates. From the Gunalan et al., 2019 (E07), results, which used a reference strain Sal1 for infections, the *pir* responses of the parasite within the four independent hosts (of two different species) are highly similar. This may be because Sal1 is a reference strain of *P. vivax* (Collins et al., 1972) and has reduced capacity for diversity of *pir* transcription as a

consequence of genetic bottlenecking from original extraction and being reared in the lab. However, it could also be because the host environment has little impact on the *pir* transcriptome. The monkeys in this study were splenectomised, and if any PIRs are involved in sequestration in that organ (or to avoid splenic clearance) I can only suppose that this could also influence the relative stability of *pir* sub-family proportions between infections. The original publication observed the upregulation of over 20 *pir*s in the *Aotus* infections, and only 3 in the *Saimiri* infections, by mapping to the Sal1 reference. This is a low proportion of the total *pir* gene repertoire of Sal1 but may also be reflecting that there are many missing many *pir*s in this reference genome. PCA of the global transcriptomes showed that *P. vivax* infections in the two different species of host are distinct from one another, however *pir*s likely do not contribute much to this difference.

From Kim et al., 2019 (E06), we can see paired timepoints of the same human infections, demonstrating how the populations at both timings of extraction have similar *pir* sub-family proportions (Figure 5.11), although some of the lower-transcribed sub-families 'drop-out' of the profile. For each of these samples mapped to the same *de novo* assembly reference it would be possible to use QUsage gene set enrichment analysis, as performed in Chapters 3 and 4, to test for statistically significant changes in sub-family transcription between timepoints/conditions (Meng et al., 2019).

Comparing the total *pir* transcriptome between pre- and post-chloroquine (CQ) patient samples (from Kim et al., 2019, E06) I observe that most of the time there is a major depletion of the multigene family after chloroquine treatment (Figure 5.11). There are three possible explanations for this response; 1) *pir* transcription is downregulated by the parasite after drug exposure; 2) CQ directly interferes with *pir* transcriptional regulation; 3) and/or that the parasites' transcriptional patterns are errant as many parasites are dying. Even if the parasite is actively regulating *pir* activity this could be necessary because *pir* expression is detrimental under CQ conditions, or because the parasite is shifting resources towards other processes. Since the *pir*s are downregulated, it is likely that their function to the parasite at this stage of development may be something the parasite can use sparingly when under duress. Notably the middle-right sample of Figure 5.11 shows an infection depleting *pir* transcription from over 2000 TPM to under 100 TPM. These parasites are dying, so even if this behaviour is unfit for the cells, it could still be seen in this population. Transcriptional changes before-and-after drug treatment have been measured multiple times in *P. falciparum* infections, and multigene

families like the *var*s/PfEMP1, *rifin*s, and *stevor*s have been among the genes showing changes between conditions, albeit with varying patterns. Surface antigens, including these *P. falciparum* multigene families, frequently have altered transcription patterns upon drug treatment (Natalang et al., 2008), an effect which may extend across species including *P. yoelii* (Silveira et al., 2007). Microarrays have shown that artemisinin treatment of trophozoites lead to enhanced transcription of four *var*s, three *rifin*s, three *stevor*s, and multiple pseudogenes of these groups, with one *var* and one *rifin* down-expressed (Shaw et al., 2015). RNAseq data of artemisinin-treated rings from the same study showed three *rifin*s to be upregulated. Proteomics after bisthiazolium T4 exposure demonstrated some PfEMP1 upregulation (le Roch et al., 2008). Evidence against the importance of these genes in drug-response phenotypes comes from the transcriptomes of *P. falciparum* isolates possessing artemisinin-resistance phenotypes. These were collected from two separate studies, and the results gave no indication of the *var*s/PfEMP1, *rifin*s, or *stevor*s, being differentially transcribed in both sets compared to susceptible parasites, beyond one upregulated *var* gene (Zhu et al., 2022). Testing a range of compounds to find general perturbation signals in the parasite found the three multigene families to be major signals (Hu et al., 2010). Hence, in *P. falciparum*, cellular stress often impacts multigene family transcription without it being a specific response to given conditions. Even taking those general changes into account, however, microarray RNA transcription measurements after exposure of *in vitro P. falciparum* to pyronaridine and chloroquine showed a significant downregulation of some *var*s (especially after chloroquine) and a significant upregulation of a handful of *rifin*s (especially after pyronaridine) (Kritsiriwuthinan et al., 2011). The relationship between multigene family transcription and treatment molecule responses is complex and warrants a systematic review or meta-analysis to clarify exactly what the literature is telling us.

Inter-species comparisons of the *pir* genes are complicated by the lack of orthologs and common families between the *P. vivax* and rodent malaria lineages. However, it is possible to find some tests we can make to investigate whether *pir* behaviour is conserved across these species. The results from *P. berghei*, *P. c. chabaudi* AS and *P. vivax* all agree that transcription in the blood is high, and the *P. berghei* and *P. vivax* data agree that mosquito and most liver-stage *pir* transcription is negligible. Transcriptional patterns across the IDC, however, do not agree. *P. c. chabaudi* AS *pir*s peak in the late trophozoite stages and perhaps early schizonts, while *P. vivax pir*s peak in the early rings and perhaps late schizont/merozoite forms. The data from *P. berghei*

collected by Otto et al., 2014, suggests that *P. berghei* and *P. vivax* both show peak transcription in rings, however the single cell data of Howick et al., 2019, when arranged by pseudotime, suggests that the trophozoites are the peak transcribers. The single-cell atlas of *P. vivax* (obtained using four different strains infecting *Aotus* and *Saimiri* monkeys) currently lacks ring stages which restricts using this resource to verify the bulk RNAseq results (Sà et al., 2020). These pieces of evidence are not ideal, as the Otto et al., 2014, evidence is based on only a few replicates and but three timepoints of the IDC, while single-cell technology often misses low transcribing genes like the *pir*s. The *in vitro P. vivax* and *P. berghei* data may also have been impacted by the toxicity of culturing conditions to both species, which could alter natural gene transcription at later time points. Ultimately for these asynchronous species scRNAseq provides the best hope of acquiring correct signals from different stages of the IDC, and as technology improves the resolution of the method will too.

Given the variation of samples there are also novel opportunities in these data to understand the dynamics of the *pir* gene family compared to laboratory models, including genetic and geographical variation of transcribed members. As well as exploring the *pir* diversity across geographic regions, as previous studies have done, one could further ascertain whether these genes are actually being transcribed. I would hypothesise that sub-families with transcription rates that vary substantially between sites may indicate groups of *pir*s which are less essential to the fundamental processes of the parasite, but which respond to environmental cues. Unfortunately, this was difficult for me to perform with the samples that have transcriptional profiles reported here. Existing Brazilian samples will provide a major new geographic region for comparison with the Thai/Cambodian samples currently included, such as those not incorporated in my analysis due to concerns over which samples to use for unique assemblies (Roth et al., 2018) and data availability at time of investigation (Bourgard et al., 2021). With a more comprehensive contingent of South American samples a comparison between continental *pir* transcription would become possible.

Chapter 6. Conclusion

Even after over a hundred years at the forefront of medical research, the disease of malaria caused by *Plasmodium spp.* parasites continues to plague low- and middle-income nations around the world. The genus includes species which invade many different classes of organism, with rodent-infecting/simian-infecting malaria parasites used as a model for the human infection. The genetics of the pathogen is now open to scientific investigation, permitting a greater understanding of how to counter the disease, but sometimes what researchers find opens a Pandora's box of further questions. Multigene families are found across the breadth of life, and the deadly malaria parasite is no exception. Identifying their purposes during the complex life cycle of *Plasmodium spp.*, interacting with both human and mosquito hosts, is a focus of modern study. The *pir* genes, perhaps the largest and most widespread gene family across the Plasmodia, has frustrated many attempts to elucidate function since their discovery by del Portillo and colleagues in 2001. Possible redundancy between the protein products in knock-out screens (Bushell et al., 2017), as well as inconsistent cellular localisation patterns between different PIRs (Fougère et al., 2016; Giorgalli et al., 2022; Yam et al., 2016), are two reasons why new results sometimes feel like they are merely muddying the water. The closest insight investigators have achieved so far is finding associations between groups of the *pir*s and virulence/chronicity. Generally, the broad transcription of many *pir*s is a signature of acute, mosquito-transmitted, low virulence infections. In contrast, limited subsets of *pir*s show higher transcription in the high-virulence blood-transmitted and chronic infections (Brugat et al., 2017; Spence et al., 2013). These studies underscored the potential importance of the genes in mediating pathophysiology, but represent only a step towards a comprehensive understanding of the family.

The thesis describes a systematic approach to analysing *pir* transcription across the life cycle of multiple species of malaria parasites (for a summary of the results see Figure 6.1). For this purpose, I used published data of rodent malaria *P. berghei* and human malaria *P. vivax,* which covered exoerythrocytic, asexual blood cycle, sexual and mosquito stages of each parasite, and the separate stages of the asexual blood cycle from novel *P. c. chabaudi* AS data. To combat issues with diversity in wild *P. vivax* isolates I applied *de novo* transcriptomics to analyse the *pir*s of this particular species. From these results, I aimed to postulate upon the function of *pir*s and *pir* sub-families by

identifying consistent patterns of gene expression associated with certain times of development. Furthermore, I discuss the conserved *pir* behaviours between rodent and human malarias, and the relevance of the former as a model for the latter. By looking across the *P. berghei* life cycle I observed peaks of *pir* transcription upon release of liver merosomes, the asexual blood cycle, and differentiation into male gametocytes. Upon fertilisation and development in the mosquito, *pir* transcription drops considerably, showing the consistency of previous results across different experiments. Although not necessarily the peaks of transcription, the asexual and exoeryrthrocytic schizonts are enriched for Long-form *pir*s, further linking this gene set to schizogony since Otto and colleagues observed the L-form association with their asexual schizonts in 2014. The related rodent malaria parasite *P. c. chabaudi* AS was used to investigate the asexual blood cycle in detail due to its synchrony. This showed high *pir* transcription across the asexual cycle but peaking during the late trophozoites and perhaps the early schizonts. As in *P. berghei*, the Long-form *pir*s are enriched in the asexual schizonts, and furthermore are transcribed across a broader set of the asexual cycle than the Short-form *pir*s. This matched the temporal distinction between ChAPL and AAPL *pir* behaviours respectively, showing that chronicity *pir*s display transcriptional differences even in acute infection. In *P. vivax* I showed that the *de novo* generation of *pir*s can produce hundreds of expressed transcripts which show sequence similarity with previously defined *pir* sub-families. Despite this, one new *pir* sub-family appeared to be highly transcribed across samples (Unassigned_3) with overall dissimilarity with reference genome *pir*s. The *pir* transcription across the intra-erythrocytic development cycle differed to that of *P. c. chabaudi* AS, with peaks during late schizonts and early ring stages. Although similarity was found between this human malaria parasite and the rodent models when it is observed that mosquito and exoerythrocytic transcription of *pir*s is minimal, and asexual transcription is high. Possible roles of Long-form *pir*s in schizogony, and of the *pir* family in asexual blood stages (but not exoerythrocytic or mosquito stages), are likely. The ancestral *pir* sequence is highly transcribed across multiple life cycle stages in different species, strongly suggesting an important role of this gene in particular.

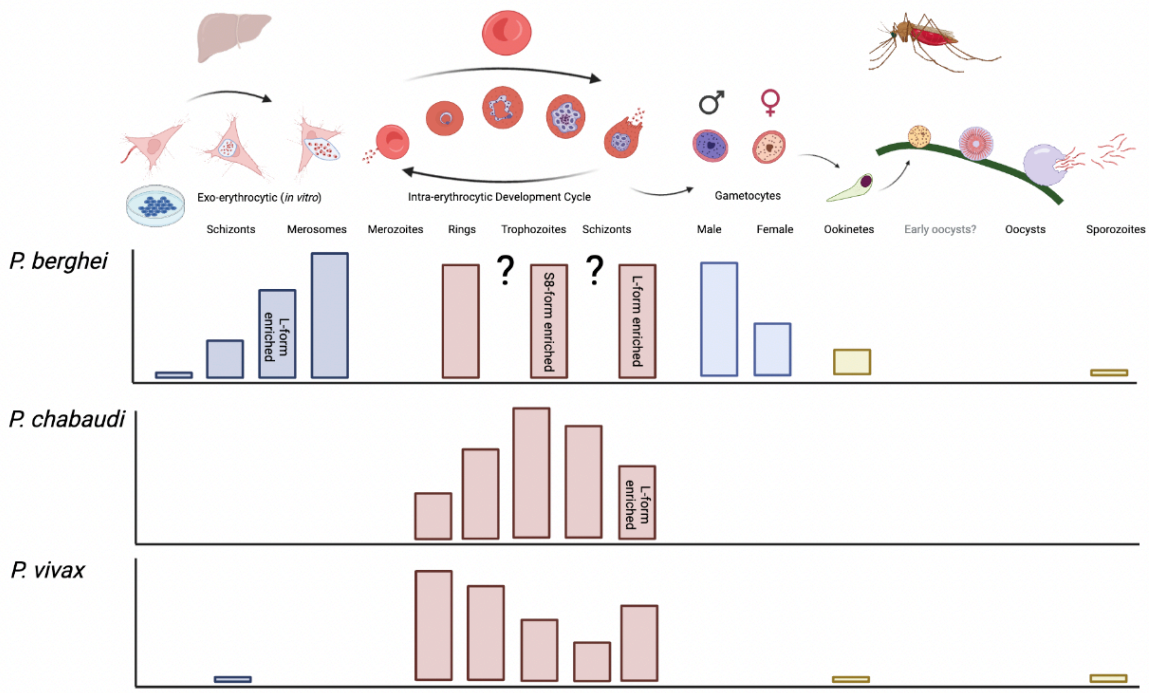


Figure 6.1 – Summary figure of *pir* transcriptional patterns observed in this thesis across the life cycle of *P. berghei*, *P. c. chabaudi* AS, and *P. vivax*.
Presence of a bar indicates that transcriptional data was available for that stage in that species and corresponds approximately to the expression level in TPM. Absence of a bar completely shows a lack of available data. Question marks highlight existing uncertainty even given the data. Within the bar are the sub-family enrichments where they have been observed. This figure also highlights the gaps in knowledge of given species as well as where existing results are in need of replication. 'Early oocysts' are a sample missing from all bulk RNAseq datasets for each species but one of particular interest. Figure drawn using Biorender (app.biorender.com).

Although many stages of the life cycle are covered it is important to remain cognizant of where the data here is from. For example, the liver stages of *P. berghei* and *P. vivax* are not from *in vivo* livers, but cell cultures, which may change transcriptional patterns of the parasites to one that they'd never demonstrate in the wild (Lapp et al., 2015). Mosquito stages are from various insect vectors, and it is unknown how transcriptional patterns are impacted by infecting different species. Comparison of RNAseq of the same parasite species infecting different mosquito species would be of general interest to many aspects of parasite-vector biology. Other missing sources of variation include that caused by the time-of-day of extraction (not just within the IDC), the length of infection

(oocysts may last for over two weeks), host species/strain, and the tissue of residence (especially for sequestering schizonts/gametocytes (Ebbinghaus & Krücken, 2011)).

Evidence of a dearth of *pir* transcription at given stages of the life cycle does not necessarily mean that the peptide products are not playing a role at that stage. *Plasmodium spp.* engage in translational regulation strategies to control their production of protein, and this occurs strongly during the gametocyte stages (Mair et al., 2006a; Yuda et al., 2015), although it occurs across the life cycle (Balu et al., 2011; Lindner et al., 2019). Male gametocytes transcribe huge quantities of *pir*, and although translational repression may not be as active in the males compared to female gametocytes (Lasonder et al., 2016; Mair et al., 2006b), mechanisms of mRNA abundance regulation still occur (Hart et al., 2019). Proteomics offers a method to circumvent the problems with the disparity between mRNA evidence and actual product levels. The low levels of PIR protein present can frustrate attempts, however prior studies have been successful in finding PIR peptides despite this (Hall et al., 2005). The evidence of PIR peptides from the multiple different proteomic studies that exist for different species would be worth collating, and perhaps re-analysing, in a similar strategy to this thesis with RNAseq datasets.

Analysing independently derived published data sets, I have confirmed/shown that mosquito stages of the *P. berghei* and *P. vivax* parasites barely transcribe *pir*. Even with *de novo* assembly, the sporozoites of *P. vivax* did not give rise to any expressed *pir* signal. Hence, this pattern may be conserved across species. This is not to say *that pir*s are completely missing; *P. berghei* ookinetes still transcribe significantly more *pir* than sporozoites, and one *P. vivax* sporozoite sample showed a tiny signal of a B_2 *pir*. This latter case may be an example of a singular divergent multigene member with a new function in a different stage (Zanghì et al., 2018), but could also just be leaky transcription or even erroneous. Immunofluorescence of five select PIRs did not show any expression in *P. berghei* day 8 or 12 oocysts, or salivary gland sporozoites (Fougère et al., 2016). These samples do not cover the full picture of mosquito development, however. The sporozoites are known to undergo physiological changes during their progression from the oocyst to host invasion (Bogale et al., 2021; Lindner et al., 2019; Roth et al., 2018; Sato et al., 2014), so the full breadth of sporozoite and other vector development stage transcription is not necessarily encapsulated by these data. To fully exclude *pir* transcription at all stages of vector development, time courses of bulk

RNAseq of the parasite from blood feed, through oocyst growth, to salivary gland sporozoites, would be worth investigation.

The liver stages of *Plasmodium spp.* are a key target of study since they represent the first growth in the mammalian host and are a source of relapsing infections from quiescent cells. When I analysed these stages of *P. berghei*, *pir* transcripts were detected; this has not been previously described from bulk RNAseq. Transcription seems to be gradual from invasion of low-*pir* sporozoites to the higher-transcribing late liver schizonts in *P. berghei*. This transcriptional explosion culminates in the liver merozoites/merosomes, which transcribe more *pir* than any other stage, in a novel phenomenon reported here. Fluorescent microscopy studies of five select PIR proteins (the same mentioned above for oocyst and sporozoite investigations) in *P. berghei* liver stages demonstrated expression of two after 40h of culture, corroborating the RNAseq evidence of later transcription in exoerythrocytic stages (Fougère et al., 2016). Unfortunately, the *P. vivax* datasets of liver stages, including the quiescent hypnozoite, did not have enough sample for the *de novo* assemblers to find *pir*s. It cannot be concluded therefore from this data whether the *P. vivax* exo-erythrocytic stages transcribe *pir* or not. The liver stages are difficult to obtain as they are usually so few in number, and they are often grown in unrealistic culturing systems to increase parasite RNA yield. Single-cell RNAseq represents the best hope for targeting these elusive stages, although bulk RNAseq could be combined with laser-capture microdissection for removing the relevant sections from infected livers (Cubi et al., 2017). Mass sporozoite infections of laboratory hosts, combined with microscopy dissection, could provide an avenue to obtain bulk RNAseq of *in vivo* liver parasites of these species and *P. c. chabaudi* AS. Using luciferase/fluorescing parasite strains may permit organ-scale imaging to narrow down highly-infected regions of the livers (Cunningham et al., 2017).

The intra-erythrocytic development cycle is potentially the most *pir*-intensive time of *Plasmodium spp*. infection. The evidence from all three species analysed in this thesis supports this assertion, with this thesis presenting novel data on the cycle of *P. c. chabaudi* AS as well as a novel *de novo* transcriptome of *P. vivax* asexual stages. When the individual cell types of the asexual blood stages are separated, I see transcriptional disparities between the species. At first glance, *P. berghei* and *P. vivax* appear to upregulate *pir*s at the late schizont/early ring stages (note that *P. vivax* has a 48h-long cycle compared to the 24h cycle of these rodent malarias). These transcriptional

changes are not statistically significant in *P. berghei*, possibly due to the low number of samples of the separate asexual stages, and they seem to disagree with the single-cell atlas (Howick et al., 2019; Otto et al., 2014). The evidence from *P. vivax* is more recent, but also based on just two replicates and cannot be verified from the existing single-cell data (Zhu et al., 2016). These conclusions need repeating with greater replicate numbers. If one or both results are reproducible, then it would present a contrast with the *P. c. chabaudi* AS *pir* transcriptome that peaks in late trophozoites and troughs during the ring stages. A comparison made across 1-1 orthologs between paired species in microarray data of *P. falciparum*, *P. knowlesi*, *P. vivax*, *P. yoelii*, *P. berghei* and *P. chabaudi* showed that genes with transcriptional variation between species were enriched among those genes primarily transcribed during the early ring and early schizont stages for individual species. Although these studies would have included few *pir*s, the expression of *pir*s coincides with genes identified by the investigations as transcriptionally variable.

Comparing between rodent- and human-infecting species of *Plasmodium* permits a verification of using mammalian models for the human disease. Similarities in broader transcription (and possibly function) are observed in the minimal mosquito and the higher asexual blood stage expression. The contrast between *P. vivax* and *P. c. chabaudi* AS IDC transcription is one observed distinction, however this needs further verification. Other preliminary results described here show that the *P, vivax pir*s may not have distinct long and short groups of *pir*s like the rodent parasites (in fact it may be more correct to say that they only have short *pir*s). Many important stages cannot be compared between *P. vivax* and *P. berghei* with existing bulk RNAseq data, such as male gametocytes and late liver schizonts. These could be performed with single-cell data, which will show further the similarities/differences between these species. Of course, making a sweeping statement over comparing the human/simian-infecting lineages and the rodent-infecting lineages would require incorporating more species from both sides. However, the bulk and single-cell RNAseq from across the life cycle are mostly unavailable for these organisms.

Separated gametocytes (and separated gametocyte sexes) were only available from *P. berghei* in this thesis. The *P. c. chabaudi* AS datasets contained negligible gametocyte levels by light microscopy, and only a small percentage by deconvolution. One *P. vivax* study, Kim et al., 2019, investigated gene clusters associated with the different

gametocyte sexes within their mixed data, but I have treated the samples just as mixed asexual stages. Confirming the conclusions of Reid et al., 2018, the bulk RNAseq of *P. berghei* gametocytes from different experiments shows that they have a consistent and massive divergence in *pir* transcription, with females not transcribing much and males transcribing a lot. Single-cell data shows that this is not the full story, as the male *pir* transcriptome is massively variable and can be both high and low (Howick et al., 2019). It would be intriguing to further interrogate the single-cell RNAseq to see if and which other genes in male gametocytes associate with high or low *pir* transcription. This may help elucidate whether either of these states are associated with aberrant development, as high proportions of male gametocytes grow defectively upon gametogenesis (Matthews et al., 2022; Sinden et al., 2010). Male gametocyte-associated proteins tend to be evolving faster than female sequences (Khan et al., 2013) - a general phenomenon seen across myriad species when male and female reproductive genes are compared (Civetta & Singh, 1995; Swanson & Vacquier, 2002) - and a multigene family would enable plenty of diversity for selection to act upon. Khan et al., 2013, further demonstrated that male-associated genes with immune epitopes are evolving particularly fast. Given the distinction between male and female gametocyte transcription, a tantalising possibility emerges of *pir*s playing a role in fertilisation. Male gametocytes transcribe some *pir*s more than the other peaks of transcription, and these could be candidates for reproductive functions. Preliminary data from *P. c. chabaudi* AS gametocytes suggests that the males also transcribe more *pir* then females, but with less of a difference than in *P. berghei*, suggesting that the role of *pir*s in gametocytes or gametes could be different between species, or indicative of sexual selection for variation (Cunningham and Reid et al., in review). Multigene families are often expressed in the sexual cells of organisms, including *rifin*s in *P. falciparum* (Petter et al., 2008), and even implicated in male reproductive fitness, such as the *6-cys* family in *P. berghei* (van Dijk et al., 2010). In multicellular organisms multigene families are involved in sperm competition (Civetta & Ranz, 2019), for example, the *male secreted short* (*mss*) genes in *Caenorhabditis briggsae* (Yin et al., 2018), and the *sperm-specific dynein intermediate chain* (*Sdic*) genes in *Drosophila melanogaster* (Nurminsky et al., 1998). The *Sdic* family members are arranged adjacent to one another on a tandem gene locus, with variable copy numbers between strains, and are considered a ‘recent’ gene family encoding a sperm motility protein (Clifton et al., 2020; Nurminsky et al., 1998; Ponce & Hartl, 2006). Despite being relatively new, these genes confer sexual advantage versus knock-outs (Yeh et al., 2012). In another similarity with *pir*s, the major

transcribed gene of the *Sdic* family is the one considered to be the original/ancestral copy (Ponce & Hartl, 2006). It may be noteworthy also that seminal proteins interact with the female immune system to prevent destruction of spermatozoa and the embryo (Kelly, 1995; S. A. Robertson et al., 2009), while *Plasmodium* male gametocytes/gametes need to avoid immune destruction in the circulation and blood meal. Male reproductive proteins often have pleiotropic functions, which would explain why transcription of fertilisation genes could occur outside of gametocytes (Poiani, 2006). When *pir* KOs are generated it would be important to assess whether they have deficiencies in fertilisation and/or transmission, including in competition assays with wild-type gametocytes.

Transcriptional changes across the life cycle of the malaria parasites are not distributed evenly between the sub-families. This is exemplified by patterns first observed in this thesis. Most notably the likely enrichment of S8 *pir*s in the asexual blood cycle of *P. berghei*, the broader peak of transcription of L-form *pir*s in the *P. c. chabaudi* AS asexual blood cycle, and the loss of I-family *pir*s in late trophozoite/early schizont transcriptomes of *P. vivax*. These groupings were defined by similarity/phylogeny, so it is interesting to see that they likely correspond to distinct functional units too. The AAPLs and ChAPLs of *P. c. chabaudi* AS also showed different transcription patterns over the IDC even though these loci were originally identified using transcription from chronic infections, which were not analysed in this thesis. The method of transmission (with a corresponding change of virulence) from natural infection via mosquito bite to blood passaging impacted both the absolute amount of *pir* transcription of *P. c. chabaudi* AS, as previously known, and how the sub-families changed over the IDC, suggesting that the transcriptional patterns can adapt to conditions. However, this demonstrates that blood-passaged samples are insufficient models for understanding natural conditions of parasite transcription. Otto et al., 2014, first suggested that the schizont stages may show preferential Long-form transcription in *P. berghei*. Comparing within and between species demonstrates that the schizont stages of both the rodent malarias are enriched for Long-form *pir*s. In *P. berghei* this is seen for both the asexual and the exoerythrocytic schizonts, but not necessarily the same members; the later liver schizonts are enriched for L1s, while asexual schizonts are enriched for L2s. Unlike in *P. berghei*, the male gametocytes of *P. c. chabaudi* AS are also enriched for Long-form *pir* transcription, so in sexual stages the L-forms may have divergent behaviour across the rodent malaria parasites (Cunningham and Reid et al., in review). The *P. vivax* sub-families do not

show the same bimodal distribution of average lengths that the rodent clades do, which means that there is currently no evidence of longer or shorter sequence lineages, or that the schizonts in this species prefer longer PIRs. However, the current pool of *de novo pir*s should be tested for fragmented genes by identifying intact N- and C-terminal domains.

The coincidence of *pir* transcription with stages preceding mass cell production has led me to hypothesise that some *pir*s are involved in cell replication or the functions of nascent cells. This matches the high *pir* transcription of *P. berghei* male gametocytes, late liver schizonts, and asexual stages, of *P. c. chabaudi* AS late trophozoite stages, and *P. vivax* late schizont stages. This hypothesis is disputed by the possible timing of *P. berghei pir* upregulation (in the rings) and the lack of *P. vivax* oocyst transcription (although this may be about the time of oocyst development analysed, as shown by the single-cell atlas of Howick et al., 2019, which suggests that early oocysts are major *pir* transcribers). Some of these discrepancies may be explained by the nature or purpose of the cell division at these times. Liver and erythrocytic schizogony both produce merozoites which need to invade a new erythrocyte, which is a function that L-form *pir*s, enriched in these stages in rodent malarias, may perform. Male gametogenesis takes place much faster than schizogony and sporogony, and the resulting gametes need to bind and fuse with female gametes but not invade any host cells. If oocyst mitosis does not require *pir*s when other stages do, this could be because of differences in sporogony compared to schizogony, or that sporozoites do not need to invade cells until after they have exited the vector. With PIR-specific antibodies it would be possible to view the localisation of the proteins during schizogony, and it would be of interest to see where they move during the process and if this differs between sub-families.

P. vivax offered the opportunity to compare *pir* sub-family diversity between different isolates of the parasite. This showed that within the same infection by the same strain, the sub-family proportions tended to remain stable. The constitution of the sub-families varied across different isolates, although the larger sub-families are almost always present, highlighting constrained diversity of transcriptional patterns. This corroborates evidence from genomes of *P. vivax* sequenced globally which have shown strong overlaps of *pir* gene sub-families (Auburn et al., 2016; Brashear et al., 2020; Chen et al., 2017; Merino et al., 2006; Minassian et al., 2021). Clusters not defined by the sub-families of del Portillo et al., 2001, and Lopez et al., 2013, were also found. However,

none of these were completely unique to the *de novo* transcripts in this thesis as they also contained sequences from the P01 and/or Sal1 genomes. One cluster identified, Unassigned_3, was notably under-represented in the Papua New Guinean P01 and Salvadorian Sal1 genomes. This may represent a geographically-enriched sub-family that is also relatively highly transcribed in these mostly Thai/Cambodian datasets. Future work could incorporate additional genomic *pir* sequences from the existing genomes of Thai and Chinese/Burmese isolates for comparison, and to test if Unassigned_3 is enriched in South-East Asian samples (Auburn et al., 2016; Chen et al., 2017; Minassian et al., 2021). Much like the proportions of the numbers of sequences, the transcriptional profiles themselves are variable but generally show the same major sub-families. Sub-family E is the largest family across assemblies, and this is reflected in the transcriptional data. It would be of interest to obtain RNAseq data from China-Myanmar border cases of *P. vivax*, as genome assemblies suggest that the C sub-family among isolates there is nearly as large as the E sub-family (note that the C sub-family family is split into two in this thesis as this better matched the similarity network), so perhaps the C sub-family will be both genomically and transcriptionally enriched in the upper Mekong region (Brashear et al., 2020).

The ancestral *pir* was highlighted as highly conserved across *Plasmodium* species by Frech and Chen et al., 2013. I observed from the transcriptional data of *P. berghei* and *P. c. chabaudi* AS that it is also one of the most consistently and highly transcribed *pir*s. In a field of generally low transcribing members of the multigene family, it stood out. Expanding my focus to virtually all known species with *pir* genes, I have evidence that this gene is often highly transcribed and hence probably plays a conserved, important role across the parasites. This extends to *P. vivax*, although not all the *de novo* transcriptome assemblies made contained this *pir*. I suggest that this is due to the fidelity of the assembly software as opposed to a lack of ancestral *pir* transcription, as *P. c. chabaudi* assemblies with known samples also did not always construct it. The next step would be to perform genetic knock-outs in rodent models to see if it genuinely does play an essential role in *Plasmodium* biology, or if there is a growth phenotype. Even if neither of these are observed, RNAseq of the KO may show that other *pir*s are upregulated to compensate for the loss of one member.

Collating existing and novel data of *pir* transcription has identified consistent patterns across the *P. berghei* life cycle, the *P. c. chabaudi* AS intra-erythrocytic development

cycle, and the *P. vivax* life cycle and IDC. Comparing between these species finds aspects of agreement and disparity of *pir* transcription, highlighting conserved behaviours as well as divergent ones. While there is no evidence of substantial *pir* transcription in the mosquito, the asexual blood stages show patterns of enhanced transcription of *pir*s. The IDC may show divergent *pir* behaviour between species. The role and localisation of PIRs in late liver schizonts, liver merosomes and male gametocytes are all exciting focusses for future study. Each of these transcriptional signals should be investigated in the species in which they are missing or in need of replication. The ancestral *pir* was found to be one that was highly expressed in most samples and species, hinting at an important and conserved function. These results can guide future targeted gene studies and create hypotheses about the functions *pir*s may or may not have, such as the putative role of Long-form *pir*s in schizogony or invasion.

Chapter 7. Reference List

Allison, A. C. (1954). Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *British Medical Journal*, *1*(4857), 290. https://doi.org/10.1136/BMJ.1.4857.290

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amin, S., Prentis, P. J., Gilding, E. K., & Pavasovic, A. (2014). Assembly and annotation of a non-model gastropod (Nerita melanotragus) transcriptome: A comparison of de novo assemblers. *BMC Research Notes*, *7*(1), 1–8. https://doi.org/10.1186/1756-0500-7-488

Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E. Y., Bażant, W., Belnap, R., Blevins, A. S., Böhme, U., Brestelli, J., Brunk, B. P., Caddick, M., Callan, D., Campbell, L., Christensen, M. B., Christophides, G. K., Crouch, K., … Zheng, J. (2022). VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Research*, *50*(D1), D898–D911. https://doi.org/10.1093/NAR/GKAB929

Andrews, S. (n.d.). *FastQC A Quality Control tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aresta-Branco, F., Sanches-Vaz, M., Bento, F., Rodrigues, J. A., & Figueiredo, L. M. (2019). African trypanosomes expressing multiple VSGs are rapidly eliminated by the host immune system. *Proceedings of the National Academy of Sciences*, *116*(41), 20725–20735. https://doi.org/10.1073/pnas.1905120116

Auburn, S., Böhme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C. I., Berriman, M., Price, R. N., & Otto, T. D. (2016). A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of *pir* genes. *Wellcome Open Research*, *1*, 4. https://doi.org/10.12688/wellcomeopenres.9876.1

Aunin, E., Böhme, U., Sanderson, T., Simons, N. D., Goldberg, T. L., Ting, N., Chapman, C. A., Newbold, C. I., Berriman, M., & Reid, A. J. (2020). Genomic and transcriptomic evidence for descent from *Plasmodium* and loss of blood schizogony

in *Hepatocystis* parasites from naturally infected red colobus monkeys. *PLoS Pathogens*, *16*(8), e1008717. https://doi.org/10.1371/JOURNAL.PPAT.1008717

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J. A., … Wang, H. (2009). PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Research*, *37*(SUPPL. 1), 539–543. https://doi.org/10.1093/nar/gkn814

Babb, P. L., Fernandez-Duque, E., Baiduc, C. A., Gagneux, P., Evans, S., & Schurr, T. G. (2011). mtDNA diversity in azara's owl monkeys (*Aotus azarai azarai*) of the Argentinean Chaco. *American Journal of Physical Anthropology*, *146*(2), 209–224. https://doi.org/10.1002/AJPA.21567

Bachmann, A., Petter, M., Krumkamp, R., Esen, M., Held, J., Scholz, J. A. M., Li, T., Sim, B. K. L., Hoffman, S. L., Kremsner, P. G., Mordmüller, B., Duffy, M. F., & Tannich, E. (2016). Mosquito Passage Dramatically Changes *var* Gene Expression in Controlled Human *Plasmodium falciparum* Infections. *PLOS Pathogens*, *12*(4), e1005538. https://doi.org/10.1371/journal.ppat.1005538

Bafort, J. (1968). Primary exo-erythrocytic forms of *Plasmodium vinckei*. *Nature*, *217*(5135), 1264–1265. https://doi.org/10.1038/2171264B0

Baird, J. K. (2009). Resistance to therapies for infection by Plasmodium vivax. In *Clinical Microbiology Reviews* (Vol. 22, Issue 3, pp. 508–534). American Society for Microbiology (ASM). https://doi.org/10.1128/CMR.00008-09

Baird, J. K. (2022). African *Plasmodium vivax* malaria improbably rare or benign. *Trends in Parasitology*, *0*(0). https://doi.org/10.1016/J.PT.2022.05.006

Balu, B., Maher, S. P., Pance, A., Chauhan, C., Naumov, A. v., Andrews, R. M., Ellis, P. D., Khan, S. M., Lin, J. wen, Janse, C. J., Rayner, J. C., & Adams, J. H. (2011). CCR4-associated factor 1 coordinates the expression of *Plasmodium falciparum* egress and invasion proteins. *Eukaryotic Cell*, *10*(9), 1257–1263. https://doi.org/10.1128/EC.05099-11

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. v., Sirotkin, A. v., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barcons-Simon, A., Cordon-Obras, C., Guizetti, J., Bryant, J. M., & Scherf, A. (2020). CRISPR Interference of a Clonally Variant GC-Rich Noncoding RNA Family Leads

to General Repression of *var* Genes in *Plasmodium falciparum*. *MBio*, *11*(1). https://doi.org/10.1128/mBio.03054-19

Barnwell, J. W., Howard, R. J., Coon, H. G., & Miller, L. H. (1983). Splenic requirement for antigenic variation and expression of the variant antigen on the erythrocyte membrane in cloned *Plasmodium knowlesi* malaria. *Infection and Immunity*, *40*(3), 985–994. http://www.ncbi.nlm.nih.gov/pubmed/6189787

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Bateman, A., Chothia, C., & Eddy, S. R. (1996). Members of the immunoglobulin superfamily in bacteria. *Protein Science*, *5*(9), 1939–1941. https://doi.org/10.1002/pro.5560050923

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. In *Source: Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 57, Issue 1).

Bernabeu, M., Lopez, F. J., Ferrer, M., Martin-Jaular, L., Razaname, A., Corradin, G., Maier, A. G., del Portillo, H. A., & Fernandez-Becerra, C. (2012). Functional analysis of *Plasmodium vivax* VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. *Cellular Microbiology*, *14*(3), 386–400. https://doi.org/10.1111/j.1462-5822.2011.01726.x

Billker, O., Lindo, V., Panico, M., Etienne, A. E., Paxton, T., Dell, A., Rogers, M., Sinden, R. E., & Morris, H. R. (1998). Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature*, *392*(6673), 289–292. https://doi.org/10.1038/32667

Bogale, H. N., Pascini, T. v., Kanatani, S., Sá, J. M., Wellems, T. E., Sinnis, P., Vega-Rodríguez, J., & Serre, D. (2021). Transcriptional heterogeneity and tightly regulated changes in gene expression during *Plasmodium berghei* sporozoite development. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(10). https://doi.org/10.1073/PNAS.2023438118/-/DCSUPPLEMENTAL

Böhme, U., Otto, T. D., Cotton, J. A., Steinbiss, S., Sanders, M., Oyola, S. O., Nicot, A., Gandon, S., Patra, K. P., Herd, C., Bushell, E., Modrzynska, K. K., Billker, O., Vinetz, J. M., Rivero, A., Newbold, C. I., & Berriman, M. (2018). Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution

in birds and mammals. *Genome Research*, *28*(4), 547–560. https://doi.org/10.1101/gr.218123.116

Boonkaew, T., Mongkol, W., Prasert, S., Paochan, P., Yoneda, S., Nguitragool, W., Kumpitak, C., Sattabongkot, J., & Kubera, A. (2020). Transcriptome analysis of *Anopheles dirus* and *Plasmodium vivax* at ookinete and oocyst stages. *Acta Tropica*, *207*, 105502. https://doi.org/10.1016/j.actatropica.2020.105502

Bourgard, C., Lopes, S. C. P., Lacerda, M. V. G., Albrecht, L., & Costa, F. T. M. (2021). A suitable RNA preparation methodology for whole transcriptome shotgun sequencing harvested from *Plasmodium vivax*-infected patients. *Scientific Reports*, *11*(1), 5089. https://doi.org/10.1038/s41598-021-84607-w

Boyd, M. F., & Kitchen, S. F. (1939). The Demonstration of Sporozoites in Human Tissues. *The American Journal of Tropical Medicine and Hygiene*, *s1-19*(1), 27–31. https://doi.org/10.4269/AJTMH.1939.S1-19.27

Boyd, M. F., & Stratman-thomas, W. K. (1933). Studies on benign tertian malaria: 4. ON the REFRACTORINESS of NEGROES to INOCULATION with *PLASMODIUM vivax*. *American Journal of Epidemiology*, *18*(2), 485–489. https://doi.org/10.1093/OXFORDJOURNALS.AJE.A117964/2/18-2-485.PDF.GIF

Bozdech, Z., Llinás, 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*(1), e5. https://doi.org/10.1371/journal.pbio.0000005

Bozdech, Z., Mok, S., Hu, G., Imwong, M., Jaidee, A., Russell, B., Ginsburg, H., Nosten, F., Day, N. P. J., White, N. J., Carlton, J. M., & Preiser, P. R. (2008). The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proceedings of the National Academy of Sciences*, *105*(42), 16290–16295. https://doi.org/10.1073/pnas.0807404105

Brashear, A. M., Huckaby, A. C., Fan, Q., Dillard, L. J., Hu, Y., Li, Y., Zhao, Y., Wang, Z., Cao, Y., Miao, J., Guler, J. L., & Cui, L. (2020). New *Plasmodium vivax* Genomes From the China-Myanmar Border. *Frontiers in Microbiology*, *0*, 1930. https://doi.org/10.3389/FMICB.2020.01930

Brashear, A. M., Roobsoong, W., Siddiqui, F. A., Nguitragool, W., Sattabongkot, J., López-Uribe, M. M., Miao, J., & Cui, L. (2019). A glance of the blood stage transcriptome of a Southeast Asian *Plasmodium ovale* isolate. *PLOS Neglected Tropical Diseases*, *13*(11), e0007850. https://doi.org/10.1371/journal.pntd.0007850

Brown, D. D., Wensink, P. C., & Jordan, E. (1972). A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *Journal of Molecular Biology*, *63*(1), 57–73. https://doi.org/10.1016/0022-2836(72)90521-9

Brugat, T., Cunningham, D., Sodenkamp, J., Coomes, S., Wilson, M., Spence, P. J., Jarra, W., Thompson, J., Scudamore, C., & Langhorne, J. (2014). Sequestration and histopathology in *Plasmodium chabaudi* malaria are influenced by the immune response in an organ-specific manner. *Cellular Microbiology*, *16*(5), 687–700. https://doi.org/10.1111/cmi.12212

Brugat, T., Reid, A. J., Lin, J., Cunningham, D., Tumwine, I., Kushinga, G., McLaughlin, S., Spence, P., Böhme, U., Sanders, M., Conteh, S., Bushell, E., Metcalf, T., Billker, O., Duffy, P. E., Newbold, C., Berriman, M., & Langhorne, J. (2017). Antibody-independent mechanisms regulate the establishment of chronic *Plasmodium* infection. *Nature Microbiology*, *2*(4), 16276. https://doi.org/10.1038/nmicrobiol.2016.276

Bushell, E., Gomes, A. R., Sanderson, T., Anar, B., Girling, G., Herd, C., Metcalf, T., Modrzynska, K., Schwach, F., Martin, R. E., Mather, M. W., McFadden, G. I., Parts, L., Rutledge, G. G., Vaidya, A. B., Wengelnik, K., Rayner, J. C., & Billker, O. (2017). Functional Profiling of a *Plasmodium* Genome Reveals an Abundance of Essential Genes. *Cell*, *170*(2), 260-272.e8. https://doi.org/10.1016/j.cell.2017.06.030

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), 1–13. https://doi.org/10.1093/GIGASCIENCE/GIZ100

Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., & Prjibelski, A. D. (2016). rnaQUAST: a quality assessment tool for *de novo* transcriptome assemblies. *Bioinformatics*, *32*(14), 2210–2212. https://doi.org/10.1093/BIOINFORMATICS/BTW218

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE*, *12*(10), e0185056. https://doi.org/10.1371/JOURNAL.PONE.0185056

Caldelari, R., Dogga, S., Schmid, M. W., Franke-Fayard, B., Janse, C. J., Soldati-Favre, D., & Heussler, V. (2019). Transcriptome analysis of *Plasmodium berghei* during exo-erythrocytic development. *Malaria Journal*, *18*(1), 330. https://doi.org/10.1186/s12936-019-2968-7

Calvocalle, J. M., Moreno, A., Eling, W. M. C., & Nardin, E. H. (1994). *In Vitro* Development of Infectious Liver Stages of *P. yoelii* and *P. berghei* Malaria in Human Cell Lines. *Experimental Parasitology*, *79*(3), 362–373. https://doi.org/10.1006/EXPR.1994.1098

Cambie, G., Landau, I., & Chabaud, A. G. (1990). *Niches horaires des trois espèces de Plasmodies coexistant chez un Rongeur de Centrafrique*. *Comptes Rendus de l'Academie Des Sciences. Serie III, Sciences de La Vie*, *310*(5), 183–188. http://www.ncbi.nlm.nih.gov/pubmed/2107960

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. R., Crabb, B. S., del Portillo, H. A., Essien, K., Feldblyum, T. v., Fernandez-Becerra, C., … Fraser-Liggett, C. M. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, *455*(7214), 757–763. https://doi.org/10.1038/nature07327

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C. L., & Huang, X. (2015). Bridger: A new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, *16*(1), 1–10. https://doi.org/10.1186/S13059-015-0596-2/FIGURES/7

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/BIOINFORMATICS/BTY560

Chen, S.-B. B., Wang, Y., Kassegne, K., Xu, B., Shen, H.-M. M., & Chen, J.-H. H. (2017). Whole-genome sequencing of a *Plasmodium vivax* clinical isolate exhibits geographical characteristics and high genetic variation in China-Myanmar border area. *BMC Genomics*, *18*(1), 131. https://doi.org/10.1186/s12864-017-3523-y

Cheng, C. W., Jongwutiwes, S., Putaporntip, C., & Jackson, A. P. (2019). Clinical expression and antigenic profiles of a *Plasmodium vivax* vaccine candidate: merozoite surface protein 7 (PvMSP-7). *Malaria Journal*, *18*(1), 197. https://doi.org/10.1186/s12936-019-2826-7

Chiou, K. L., Pozzi, L., Lynch Alfaro, J. W., & di Fiore, A. (2011). Pleistocene diversification of living squirrel monkeys (*Saimiri spp.*) inferred from complete mitochondrial genome sequences. *Molecular Phylogenetics and Evolution*, *59*(3), 736–745. https://doi.org/10.1016/J.YMPEV.2011.03.025

Chopra, R., Burow, G., Farmer, A., Mudge, J., Simpson, C. E., & Burow, M. D. (2014). Comparisons of de novo transcriptome assemblers in diploid and polyploid species

using peanut (Arachis spp.) RNA-Seq data. *PLoS ONE*, *9*(12), e115055. https://doi.org/10.1371/journal.pone.0115055

Civetta, A., & Ranz, J. M. (2019). Genetic Factors Influencing Sperm Competition. *Frontiers in Genetics*, *10*, 820. https://doi.org/10.3389/FGENE.2019.00820/BIBTEX

Civetta, A., & Singh, R. S. (1995). High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *Journal of Molecular Evolution*, *41*(6), 1085–1095. https://doi.org/10.1007/BF00173190

Clark, H. C. (1915). THE DIAGNOSTIC VALUE OF THE PLACENTAL BLOOD FILM IN ÆSTIVO-AUTUMNAL MALARIA. *Journal of Experimental Medicine*, *22*(4), 427–444. https://doi.org/10.1084/JEM.22.4.427

Clifton, B. D., Jimenez, J., Kimura, A., Chahine, Z., Librado, P., Sanchez-Gracia, A., Abbassi, M., Carranza, F., Chan, C., Marchetti, M., Zhang, W., Shi, M., Vu, C., Yeh, S., Fanti, L., Xia, X. Q., Rozas, J., & Ranz, J. M. (2020). Understanding the Early Evolutionary Stages of a Tandem *Drosophila melanogaster*-Specific Gene Family: A Structural and Functional Population Study. *Molecular Biology and Evolution*, *37*(9), 2584–2600. https://doi.org/10.1093/MOLBEV/MSAA109

Collins, W. E., Contacos, P. G., Krotoski, W. A., & Howard, W. A. (1972). Transmission of four Central American strains of *Plasmodium vivax* from monkey to man. *The Journal of Parasitology*, *58*(2), 332–335. https://doi.org/10.2307/3278097

Collins, W. E., Sullivan, J. S., Strobert, E., Galland, G. G., Williams, A., Nace, D., Williams, T., & Barnwell, J. W. (2009). Studies on the Salvador I Strain of Plasmodium vivax in Non-human Primates and Anopheline Mosquitoes. *The American Journal of Tropical Medicine and Hygiene*, *80*(2), 228–235. https://doi.org/10.4269/AJTMH.2009.80.228

Cordy, R. J., Patrapuvich, R., Lili, L. N., Cabrera-Mora, M., Chien, J.-T., Tharp, G. K., Khadka, M., Meyer, E. V., Lapp, S. A., Joyner, C. J., Garcia, A., Banton, S., Tran, V., Luvira, V., Rungin, S., Saeseu, T., Rachaphaew, N., … Galinski, M. R. (2019). Distinct amino acid and lipid perturbations characterize acute versus chronic malaria. *JCI Insight*, *4*(9). https://doi.org/10.1172/jci.insight.125156

Cowman, A. F., & Lew, A. M. (1989). Antifolate drug selection results in duplication and rearrangement of chromosome 7 in *Plasmodium chabaudi*. *Molecular and Cellular Biology*, *9*(11), 5182–5188. https://doi.org/10.1128/MCB.9.11.5182-5188.1989

Cox, J., Semoff, S., & Hommel, M. (1987). *Plasmodium chabaudi*: a rodent malaria model for *in-vivo* and *in-vitro* cytoadherence of malaria parasites in the absence of

knobs. *Parasite Immunology*, *9*(5), 543–561. https://doi.org/10.1111/j.1365-3024.1987.tb00529.x

Cubi, R., Vembar, S. S., Biton, A., Franetich, J.-F. F., Bordessoulles, M., Sossau, D., Zanghi, G., Bosson-Vanga, H., Benard, M., Moreno, A., Dereuddre-Bosquet, N., le Grand, R., Scherf, A., & Mazier, D. (2017). Laser capture microdissection enables transcriptomic analysis of dividing and quiescent liver stages of *Plasmodium* relapsing species. *Cellular Microbiology*, *19*(8), 1–9. https://doi.org/10.1111/cmi.12735

Culleton, R., Ndounga, M., Zeyrek, F. Y., Coban, C., Casimiro, P. N., Takeo, S., Tsuboi, T., Yadava, A., Carter, R., & Tanabe, K. (2009). Evidence for the Transmission of Plasmodium vivax in the Republic of the Congo, West Central Africa. *The Journal of Infectious Diseases*, *200*(9), 1465–1469. https://doi.org/10.1086/644510

Cunningham, D. A., Jarra, W., Koernig, S., Fonager, J., Fernandez-Reyes, D., Blythe, J. E., Waller, C., Preiser, P. R., & Langhorne, J. (2005). Host immunity modulates transcriptional changes in a multigene family (*yir*) of rodent malaria. *Molecular Microbiology*, *58*(3), 636–647. https://doi.org/10.1111/j.1365-2958.2005.04840.x

Cunningham, D. A., Lin, J., Brugat, T., Jarra, W., Tumwine, I., Kushinga, G., Ramesar, J., Franke-Fayard, B., & Langhorne, J. (2017). ICAM-1 is a key receptor mediating cytoadherence and pathology in the *Plasmodium chabaudi* malaria model. *Malaria Journal*, *16*(1), 185. https://doi.org/10.1186/s12936-017-1834-8

Cunningham, D., Fonager, J., Jarra, W., Carret, C., Preiser, P., & Langhorne, J. (2009). Rapid changes in transcription profiles of the *Plasmodium yoelii yir* multigene family in clonal populations: Lack of epigenetic memory? *PLoS ONE*, *4*(1), 19–22. https://doi.org/10.1371/journal.pone.0004285

Cunningham, D., Lawton, J., Jarra, W., Preiser, P., & Langhorne, J. (2010). The *pir* multigene family of *Plasmodium*: Antigenic variation and beyond. *Molecular and Biochemical Parasitology*, *170*(2), 65–73. https://doi.org/10.1016/j.molbiopara.2009.12.010

de Vries, A., & Ripley, B. D. (2020). *ggdendro: Create Dendrograms and Tree Diagrams Using "ggplot2"* (0.1.22).

del Portillo, H. A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C. P., Schneider, N. K., Villalobos, J. M., Rajandream, M. A., Harris, D., da Silva, L. H. P., Barrell, B., & Lanzer, M. (2001). A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature*, *410*(6830), 839–842. https://doi.org/10.1038/35071118

Dessens, J. T., Sidén-Kiamos, I., Mendoza, J., Mahairaki, V., Khater, E., Vlachou, D., Xu, X. J., Kafatos, F. C., Louis, C., Dimopoulos, G., & Sinden, R. E. (2003). Soap, a novel malaria ookinete protein involved in mosquito midgut invasion and oocyst development. *Molecular Microbiology*, *49*(2), 319–329. https://doi.org/10.1046/j.1365-2958.2003.03566.x

di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. In *Nature Biotechnology* (Vol. 35, Issue 4, pp. 316–319). Nature Publishing Group. https://doi.org/10.1038/nbt.3820

Dorin, D., Alano, P., Boccaccio, I., Cicéron, L., Doerig, C., Sulpice, R., Parzy, D., & Doerig, C. (1999). An Atypical Mitogen-activated Protein Kinase (MAPK) Homologue Expressed in Gametocytes of the Human Malaria Parasite *Plasmodium falciparum*: IDENTIFICATION OF A MAPK SIGNATURE. *Journal of Biological Chemistry*, *274*(42), 29912–29920. https://doi.org/10.1074/JBC.274.42.29912

Dowle, M., & Srinivasan, A. (2019). *data.table: Extension of `data.frame`. R package version* (1.12.8). https://cran.r-project.org/package=data.table

Ebbinghaus, P., & Krücken, J. (2011). Characterization and tissue-specific expression patterns of the *Plasmodium chabaudi cir* multigene family. *Malaria Journal*, *10*(1), 1–19. https://doi.org/10.1186/1475-2875-10-272

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), 1002195. https://doi.org/10.1371/JOURNAL.PCBI.1002195

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. In *Nature Biotechnology* (Vol. 38, Issue 3, pp. 276–278). Nature Research. https://doi.org/10.1038/s41587-020-0439-x

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fernandez-Becerra, C., Bernabeu, M., Castellanos, A., Correa, B. R., Obadia, T., Ramirez, M., Rui, E., Hentzschel, F., López-Montañés, M., Ayllon-Hermida, A., Martin-Jaular, L., Elizalde-Torrent, A., Siba, P., Vêncio, R. Z., Arevalo-Herrera, M., Herrera, S., Alonso, P. L., … del Portillo, H. A. (2020). *Plasmodium vivax* spleen-

dependent genes encode antigens associated with cytoadhesion and clinical protection. *Proceedings of the National Academy of Sciences*, 201920596. https://doi.org/10.1073/pnas.1920596117

Flannery, E. L., Kangwanrangsan, N., Chuenchob, V., Roobsoong, W., Fishbaugher, M., Zhou, K., Billman, Z. P., Martinson, T., Olsen, T. M., Schäfer, C., Campo, B., Murphy, S. C., Mikolajczak, S. A., Kappe, S. H. I., & Sattabongkot, J. (2022). Plasmodium vivax latent liver infection is characterized by persistent hypnozoites, hypnozoite-derived schizonts, and time-dependent efficacy of primaquine. *Molecular Therapy - Methods and Clinical Development*, *26*, 427–440. https://doi.org/10.1016/j.omtm.2022.07.016

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*. *Molecular and Cellular Proteomics*, *10*(8). https://doi.org/10.1074/mcp.M110.006411

Fougère, A., Jackson, A. P., Paraskevi Bechtsi, D., Braks, J. A. M., Annoura, T., Fonager, J., Spaccapelo, R., Ramesar, J., Chevalley-Maurel, S., Klop, O., van der Laan, A. M. A., Tanke, H. J., Kocken, C. H. M., Pasini, E. M., Khan, S. M., Böhme, U., van Ooij, C., … Franke-Fayard, B. (2016). Variant Exported Blood-Stage Proteins Encoded by *Plasmodium* Multigene Families Are Expressed in Liver Stages Where They Are Exported into the Parasitophorous Vacuole. *PLoS Pathogens*, *12*(11), 1–37. https://doi.org/10.1371/journal.ppat.1005917

Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C., & D Haddock, S. H. (2013). A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics*, *14*(1), 1–12. https://doi.org/10.1186/1471-2164-14-167

Frech, C., & Chen, N. (2013). Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. *BMC Genomics*, *14*(1), 427. https://doi.org/10.1186/1471-2164-14-427

Fried, M., & Duffy, P. E. (1996). Adherence of *Plasmodium falciparum* to Chondroitin Sulfate A in the Human Placenta. *Science*, *272*(5267), 1502–1504. https://doi.org/10.1126/SCIENCE.272.5267.1502

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/BIOINFORMATICS/BTS565

Galen, S. C., Borner, J., Martinsen, E. S., Schaer, J., Austin, C. C., West, C. J., & Perkins, S. L. (2018). The polyphyly of *Plasmodium*: comprehensive phylogenetic analyses of the malaria parasites (order Haemosporida) reveal widespread taxonomic conflict. *Royal Society Open Science*, *5*(5), 171780. https://doi.org/10.1098/rsos.171780

Garnier, S. (2018). *viridis: Default Color Maps from "matplotlib". R package version 0.5.1*.

Gilbert, D. (2016a). Accurate & complete gene construction with EvidentialGene . *F1000Research*, *5*. https://doi.org/10.7490/F1000RESEARCH.1112467.1

Gilbert, D. (2016b). Perfect Arthropod Genes constructed with Gigabases of RNA . *F1000Research*, *5*. https://doi.org/10.7490/F1000RESEARCH.1112595.1

Gilbert, D. G. (2019a). Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene. *PeerJ*, *7*(2), e6374. https://doi.org/10.7717/peerj.6374

Gilbert, D. G. (2019b). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? *BioRxiv*, 829184. https://doi.org/10.1101/829184

Giorgalli, M., Cunningham, D. A., Broncel, M., Sait, A., Harrison, T. E., Hosking, C., Vandomme, A., Amis, S. I., Antonello, A., Sullivan, L., Uwadiae, F., Torella, L., Higgins, M. K., & Langhorne, J. (2022). Differential Trafficking and Expression of PIR Proteins in Acute and Chronic *Plasmodium* Infections. *Frontiers in Cellular and Infection Microbiology*, *0*, 746. https://doi.org/10.3389/FCIMB.2022.877253

Golgi, C. (1886). *Sull'Infezione malarica*. *Archivo per Le Scienze Mediche*, *10*, 109–135.

Gómez-Díaz, E., Yerbanga, R. S., Lefèvre, T., Cohuet, A., Rowley, M. J., Ouedraogo, J. B., & Corces, V. G. (2017). Epigenetic regulation of Plasmodium falciparum clonally variant gene expression during development in Anopheles gambiae. *Scientific Reports 2017 7:1*, *7*(1), 1–17. https://doi.org/10.1038/srep40655

Gouy, M., Guindon, S., & Gascuel, O. (2010). Sea view version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, *27*(2), 221–224. https://doi.org/10.1093/molbev/msp259

Grassi, B. (1901). *Studi di uno zoologo sulla malaria*. Roma : R. Accademia dei lincei. http://www.biodiversitylibrary.org/title/37999

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Gunalan, K., Sá, J. M., Moraes Barros, R. R., Anzick, S. L., Caleon, R. L., Mershon, J. P., Kanakabandi, K., Paneru, M., Virtaneva, K., Martens, C., Barnwell, J. W., Ribeiro, J. M., & Miller, L. H. (2019). Transcriptome profiling of *Plasmodium vivax* in *Saimiri* monkeys identifies potential ligands for invasion. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(14). https://doi.org/10.1073/pnas.1818485116

Gural, N., Mancio-Silva, L., Miller, A. B., Galstian, A., Butty, V. L., Levine, S. S., Patrapuvich, R., Desai, S. P., Mikolajczak, S. A., Kappe, S. H. I., Fleming, H. E., March, S., Sattabongkot, J., & Bhatia, S. N. (2018). In Vitro Culture, Drug Sensitivity, and Transcriptome of *Plasmodium vivax* Hypnozoites. *Cell Host & Microbe*, *23*(3), 395-406.e4. https://doi.org/10.1016/j.chom.2018.01.002

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Haldane, J. B. S. (1949). THE RATE OF MUTATION OF HUMAN GENES. *Hereditas*, *35*(S1), 267–273. https://doi.org/10.1111/j.1601-5223.1949.tb03339.x

Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W. A., 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., … Sinden, R. E. (2005). A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, *307*(5706), 82–86. https://doi.org/10.1126/science.1103717

Hamada, S., Pionneau, C., Parizot, C., Silvie, O., Chardonnet, S., & Marinach, C. (2021). In-depth proteomic analysis of *Plasmodium berghei* sporozoites using trapped ion mobility spectrometry with parallel accumulation-serial fragmentation. *PROTEOMICS*, *2021*, 2000305. https://doi.org/10.1002/pmic.202000305

Hamblin, M. T., & Di Rienzo, A. (2000). Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. *The American Journal of Human Genetics*, *66*(5), 1669–1679. https://doi.org/10.1086/302879

Harrison, T. E., Reid, A. J., Cunningham, D., Langhorne, J., & Higgins, M. K. (2020). Structure of the *Plasmodium interspersed repeat* proteins of the malaria parasite. *Proceedings of the National Academy of Sciences*, 202016775. https://doi.org/10.1073/pnas.2016775117

Hart, K. J., Oberstaller, J., Walker, M. P., Minns, A. M., Kennedy, M. F., Padykula, I., Adams, J. H., & Lindner, S. E. (2019). *Plasmodium* male gametocyte development and transmission are critically regulated by the two putative deadenylases of the CAF1/CCR4/NOT complex. *PLOS Pathogens*, *15*(1), e1007164. https://doi.org/10.1371/journal.ppat.1007164

Hart, T., Komori, H., LaMere, S., Podshivalova, K., & Salomon, D. R. (2013). Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics*, *14*(1), 778. https://doi.org/10.1186/1471-2164-14-778

Hawking, F., Gammage, K., & Worms, M. J. (1972). The asexual and sexual circadian rhythms of *Plasmodium vinckei chabaudi* , of *P. berghei* and of *P. gallinaceum*. *Parasitology*, *65*(2), 189–201. https://doi.org/10.1017/S003118200004498X

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., & Teichmann, S. A. (2014). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology*, *7*(1), 497–497. https://doi.org/10.1038/msb.2011.28

Hirako, I. C., Assis, P. A., Hojo-Souza, N. S., Reed, G., Nakaya, H., Golenbock, D. T., Coimbra, R. S., & Gazzinelli, R. T. (2018). Daily Rhythms of TNFα Expression and Food Intake Regulate Synchrony of Plasmodium Stages with the Host Circadian Cycle. *Cell Host & Microbe*, *0*(0). https://doi.org/10.1016/j.chom.2018.04.016

Hölzer, M., & Marz, M. (2019). *De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, *8*(5), 1–16. https://doi.org/10.1093/gigascience/giz039

Hommel, M., David, P. H., Guillotte, M., & Pereira da Silva, L. (1982). Protection against plasmodiu chabaudi malaria I. — Vaccination of mice with merozoites and Freund's

adjuvants. *Annales de l'Institut Pasteur / Immunologie*, *133*(1), 57–67. https://doi.org/10.1016/0769-2625(82)90006-X

Hoo, R., Zhu, L., Amaladoss, A., Mok, S., Natalang, O., Lapp, S. A., Hu, G., Liew, K., Galinski, M. R., Bozdech, Z., & Preiser, P. R. (2016). Integrated analysis of the *Plasmodium* species transcriptome. *EBioMedicine*, *7*, 255–266. https://doi.org/10.1016/j.ebiom.2016.04.011

Howick, V. M., Russell, A. J. C., Andrews, T., Heaton, H., Reid, A. J., Natarajan, K., Butungi, H., Metcalf, T., Verzier, L. H., Rayner, J. C., Berriman, M., Herren, J. K., Billker, O., Hemberg, M., Talman, A. M., & Lawniczak, M. K. N. (2019). The Malaria Cell Atlas: Single parasite transcriptomes across the complete *Plasmodium* life cycle. *Science*, *365*(6455), eaaw2619. https://doi.org/10.1126/SCIENCE.AAW2619

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. *Nature Biotechnology 2009 28:1*, *28*(1), 91–98. https://doi.org/10.1038/nbt.1597

Huber, M., Cabib, E., & Miller, L. H. (1991). Malaria parasite chitinase and penetration of the mosquito peritrophic membrane. *Proceedings of the National Academy of Sciences*, *88*(7), 2807–2810. https://doi.org/10.1073/pnas.88.7.2807

Hunt, P. (2007). *Poverty, Malaria and the Right to Health : Exploring the Connections*. UN Chronicle. https://www.un.org/en/chronicle/article/poverty-malaria-and-right-health-exploring-connections

Inoue, N., Ikawa, M., Isotani, A., & Okabe, M. (2005). The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs. *Nature*, *434*(7030), 234–238. https://doi.org/10.1038/nature03362

James, S. P., & Tate, P. (1938). Exo-erythrocytic schizogony in *Plasmodium gallinaceum* Brumpt, 1935. *Parasitology*, *30*(1), 128–139. https://doi.org/10.1017/S0031182000010891

Janse, C. J., & Waters, A. P. (1995). Plasmodium berghei: The application of cultivation and purification techniques to molecular studies of malaria parasites. *Parasitology Today*, *11*(4), 138–143. https://doi.org/10.1016/0169-4758(95)80133-2

Janssen, C. S., Barrett, M. P., Turner, C. M. R., Phillips, R. S., & Stephen Phillips, R. (2002). A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1489), 431–436. https://doi.org/10.1098/rspb.2001.1903

Janssen, C. S., Phillips, R. S., Turner, M. R., & Barret, M. P. (2004). *Plasmodium interspersed repeats*: The major multigene superfamily of malaria parasites. *Nucleic Acids Research*, *32*(19), 5712–5720. https://doi.org/10.1093/nar/gkh907

Jiang, R., Sun, T., Song, D., & Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology*, *23*(1). https://doi.org/10.1186/S13059-022-02601-5

Kaiser, K., Camargo, N., & Kappe, S. H. I. (2003). Transformation of sporozoites into early exoerythrocytic malaria parasites does not require host cells. *The Journal of Experimental Medicine*, *197*(8), 1045–1050. https://doi.org/10.1084/jem.20022100

Kamaliddin, C., Rombaut, D., Guillochon, E., Royo, J., Ezinmegnon, S., Agbota, G., Huguet, S., Guemouri, S., Peirera, C., Coppée, R., Broussard, C., Alao, J. M., Aubouy, A., Guillonneau, F., Deloron, P., & Bertin, G. I. (2019). From genomic to LC-MS/MS evidence: Analysis of PfEMP1 in Benin malaria cases. *PLOS ONE*, *14*(6), e0218012. https://doi.org/10.1371/journal.pone.0218012

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Kelly, R. W. (1995). Contraception: Immunosuppressive mechanisms in semen: implications for contraception. *Human Reproduction*, *10*(7), 1686–1693. https://doi.org/10.1093/OXFORDJOURNALS.HUMREP.A136156

Khan, S. M., Reece, S. E., Waters, A. P., Janse, C. J., & Kaczanowski, S. (2013). Why are male malaria parasites in such a rush?Sex-specific evolution and host–parasite interactions. *Evolution, Medicine, and Public Health*, *2013*(1), 3–13. https://doi.org/10.1093/EMPH/EOS003

Kharchenko, P. V, Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740–742. https://doi.org/10.1038/nmeth.2967

Kho, S., Qotrunnada, L., Leonardo, L., Andries, B., Wardani, P. A. I., Fricot, A., Henry, B., Hardy, D., Margyaningsih, N. I., Apriyanti, D., Puspitasari, A. M., Prayoga, P., Trianty, L., Kenangalem, E., Chretien, F., Brousse, V., Safeukui, I., … Anstey, N. M. (2021). Evaluation of splenic accumulation and colocalization of immature reticulocytes and Plasmodium vivax in asymptomatic malaria: A prospective human splenectomy study. *PLOS Medicine*, *18*(5), e1003632. https://doi.org/10.1371/journal.pmed.1003632

Killick-Kendrick, R., & Peters, W. (1978). *Rodent malaria*. Elsevier.

Kim, A., Popovici, J., Menard, D., & Serre, D. (2019). *Plasmodium vivax* transcriptomes reveal stage-specific chloroquine response and differential regulation of male and female gametocytes. *Nature Communications*, *10*(1), 371. https://doi.org/10.1038/s41467-019-08312-z

Kim, A., Popovici, J., Vantaux, A., Samreth, R., Bin, S., Kim, S., Roesch, C., Liang, L., Davies, H., Felgner, P., Herrera, S., Arévalo-Herrera, M., Ménard, D., & Serre, D. (2017). Characterization of *P. vivax* blood stage transcriptomes from field isolates reveals similarities among infections and complex gene isoforms. *Scientific Reports*, *7*(1), 7761. https://doi.org/10.1038/s41598-017-07275-9

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kitchen, S. F. (1938). The Infection of Reticulocytes by *Plasmodium vivax*. *American Journal of Tropical Medicine*, *18*(4), 347–359.

Kligler, I. J., & Mer, G. (1937). Studies on the Effect of Various Factors on the Infection Rate of Anopheles Elutus with Different Species of Plasmodium. *Annals of Tropical Medicine & Parasitology*, *31*(1), 71–83. https://doi.org/10.1080/00034983.1937.11684967

Kozak, M., & Powers, S. J. (2017). If not multiple comparisons, then what? *Annals of Applied Biology*, *171*(3), 277–280. https://doi.org/10.1111/aab.12379

Kritsiriwuthinan, K., Shaw, P. J., Wongsombat, C., Chavalitshewinkoon-Petmitr, P., & Kamchonwongpaisan, S. (2011). Global gene expression profiling of Plasmodium falciparum in response to the anti-malarial drug pyronaridine. *Malaria Journal*, *10*(1), 1–10. https://doi.org/10.1186/1475-2875-10-242/TABLES/2

Krotoski, W. A. (1985). Discovery of the hypnozoite and a new theory of malarial relapse. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *79*(1), 1. https://doi.org/10.1016/0035-9203(85)90221-4

Krotoski, W. A., Collins, W. E., Bray, R. S., Garnham, P. C., Cogswell, F. B., Gwadz, R. W., Killick-Kendrick, R., Wolf, R., Sinden, R., Koontz, L. C., & Stanfill, P. S. (1982). Demonstration of Hypnozoites in Sporozoite-Transmitted *Plasmodium vivax* Infection. *The American Journal of Tropical Medicine and Hygiene*, *31*(6), 1291–1293. https://doi.org/10.4269/AJTMH.1982.31.1291

Krotoski, W. A., Krotoski, D. M., Garnham, P. C. C., Bray, R. S., Killick-Kendrick, R., Draper, C. C., Targett, G. A. T., & Guy, M. W. (1980). Relapses in primate malaria:

discovery of two populations of exoerythrocytic stages. Preliminary note. *British Medical Journal*, *280*(6208), 153. https://doi.org/10.1136/BMJ.280.6208.153-A

Krueger, F. (2015). *Trim Galore!* http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Lambros, C., & Vanderberg, J. P. (1979). Synchronization of Plasmodium falciparum erythrocytic stages in culture. *Journal of Parasitology*, *65*(3), 418–420. https://doi.org/10.2307/3280287

Landau, I. (1965). *Description de Plasmodium chabaudi n. sp., parasite de rongeurs africains*. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, *260*, 3758–3761.

Landau, I., & Killick-Kendrick, R. (1966). Rodent plasmodia of the République Centrafricaine: The sporogony and tissue stages of *Plasmodium chabaudi* and *P. berghei yoelii*. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *60*(5), 633–649. https://doi.org/10.1016/0035-9203(66)90010-1

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., … Trust:, T. W. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Langhorne, J., Cross, C., Seixas, E., Li, C., & von der Weid, T. (1998). A role for B cells in the development of T cell helper function in a malaria infection in mice. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(4), 1730–1734. https://doi.org/10.1073/pnas.95.4.1730

Langhorne, J., Gillard, S., Simon, B., Slade, S., & Eichmann, K. (1989). Frequencies of CD4+ T cells reactive with *Plasmodium chabaudi chabaudi*: distinct response kinetics for cells with Th1 and Th2 characteristics during infection. *International Immunology*, *1*(4), 416–424. https://doi.org/10.1093/INTIMM/1.4.416

Langhorne, J., Ndungu, F. M., Sponaas, A.-M., & Marsh, K. (2008). Immunity to malaria: more questions than answers. *Nature Immunology*, *9*(7), 725–732. https://doi.org/10.1038/ni.f.205

Lansink, M., Engel, J. A., Jae Lee, H., Soon, M. S., SheelaNair, A., Pernold, C. P., Laohamonthonkul, P., Akter, J., Hill, M. M., Talman, A. M., Russell, A., Lawniczak, M., Davenport, M. P., Khoury, D. S., & Haque, A. (2021). Systemic host inflammation induces stage-specific transcriptomic modification and slower maturation in malaria parasites. *BioRxiv*. https://doi.org/10.1101/2021.08.18.456784

Lapp, S. A., Mok, S., Zhu, L., Wu, H., Preiser, P. R., Bozdech, Z., & Galinski, M. R. (2015). *Plasmodium knowlesi* gene expression differs in *ex vivo* compared to *in vitro* blood-stage cultures. *Malaria Journal*, *14*(1). https://doi.org/10.1186/s12936-015-0612-8

Lasonder, E., Rijpma, S. R., van Schaijk, B. C. L., Hoeijmakers, W. A. M., Kensche, P. R., Gresnigt, M. S., Italiaander, A., Vos, M. W., Woestenenk, R., Bousema, T., Mair, G. R., Khan, S. M., Janse, C. J., Bártfai, R., & Sauerwein, R. W. (2016). Integrated transcriptomic and proteomic analyses of *P. falciparum* gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Research*, *44*(13), 6087–6101. https://doi.org/10.1093/NAR/GKW536

Lau, C. K. Y., Turner, L., Jespersen, J. S., Lowe, E. D., Petersen, B., Wang, C. W., Petersen, J. E. V., Lusingu, J., Theander, T. G., Lavstsen, T., & Higgins, M. K. (2015). Structural Conservation Despite Huge Sequence Diversity Allows EPCR Binding by the PfEMP1 Family Implicated in Severe Childhood Malaria. *Cell Host & Microbe*, *17*(1), 118. https://doi.org/10.1016/J.CHOM.2014.11.007

Laveran, A. (1881). *Nature parasitaire des accidents de l'impaludisme: description d'un nouveau parasite trouve dans le sang des malades atteints de fievre palustre.* Bailliere.

Lawton, J., Brugat, T., Yan, Y., Reid, A., Böhme, U., Otto, T., Pain, A., Jackson, A., Berriman, M., Cunningham, D., Preiser, P., & Langhorne, J. (2012). Characterization and gene expression analysis of the *cir* multi-gene family of *Plasmodium chabaudi chabaudi* (AS). *BMC Genomics*, *13*(1), 125. https://doi.org/10.1186/1471-2164-13-125

le Roch, K. G., Johnson, J. R., Ahiboh, H., Chung, D. W. D., Prudhomme, J., Plouffe, D., Henson, K., Zhou, Y., Witola, W., Yates, J. R., Mamoun, C. ben, 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*(1), 1–15. https://doi.org/10.1186/1471-2164-9-513/FIGURES/5

LeRoux, M., Lakshmanan, V., & Daily, J. P. (2009). *Plasmodium falciparum* biology: analysis of in vitro versus in vivo growth conditions. *Trends in Parasitology*, *25*(10), 474–481. https://doi.org/10.1016/J.PT.2009.07.005

Levine, N. (1970). Taxonomy of the Sporozoa. *The Journal of Parasitology*, *56*(4 Sect. 2, Part 1: Supplement: Proceedings Of the Second International Congress of Parasitology), 208–209. https://doi.org/10.2307/3277701

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/BIOINFORMATICS/BTL158

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lin, J.-W., Reid, A. J., Cunningham, D., Bohme, U., Tumwine, I., Keller-Mclaughlin, S., Sanders, M., Berriman, M., & Langhorne, J. (2018). Genomic and transcriptomic comparisons of closely related malaria parasites differing in virulence and sequestration pattern. *Wellcome Open Research*, *3*, 142. https://doi.org/10.12688/wellcomeopenres.14797.1

Lindner, S. E., Swearingen, K. E., Shears, M. J., Walker, M. P., Vrana, E. N., Hart, K. J., Minns, A. M., Sinnis, P., Moritz, R. L., & Kappe, S. H. I. (2019). Transcriptomics and proteomics reveal two waves of translational repression during the maturation of malaria parasite sporozoites. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-12936-6

Linke, A., Kühn, R., Müller, W., Honarvar, N., Ching, L. I., & Langhorne, J. (1996). *Plasmodium chabaudi chabaudi*: differential susceptibility of gene-targeted mice deficient in IL-10 to an erythrocytic-stage infection. *Experimental Parasitology*, *84*(2), 253–263. https://doi.org/10.1006/EXPR.1996.0111

Little, T. S., Cunningham, D. A., Vandomme, A., Lopez, C. T., Amis, S., Alder, C., Addy, J. W. G., McLaughlin, S., Hosking, C., Christophides, G., Reid, A. J., & Langhorne, J. (2021). Analysis of *pir* gene expression across the *Plasmodium* life cycle. *Malaria Journal*, *20*(1), 445. https://doi.org/10.1186/s12936-021-03979-6

Liu, Y., Tewari, R., Ning, J., Blagborough, A. M., Garbom, S., Pei, J., Grishin, N. V., Steele, R. E., Sinden, R. E., Snell, W. J., & Billker, O. (2008). The conserved plant sterility gene HAP2 functions after attachment of fusogenic membranes in Chlamydomonas and Plasmodium gametes. *Genes & Development*, *22*(8), 1051–1068. https://doi.org/10.1101/GAD.1656508

Logan-Klumpler, F. J., de Silva, N., Boehme, U., Rogers, M. B., Velarde, G., McQuillan, J. A., Carver, T., Aslett, M., Olsen, C., Subramanian, S., Phan, I., Farris, C., Mitra, S., Ramasamy, G., Wang, H., Tivey, A., Jackson, A., … Berriman, M. (2012).

GeneDB--an annotation database for pathogens. *Nucleic Acids Research*, *40*(D1), D98–D108. https://doi.org/10.1093/nar/gkr1032

Lopez, F. J., Bernabeu, M., Fernandez-Becerra, C., & del Portillo, H. A. (2013). A new computational approach redefines the subtelomeric *vir* superfamily of *Plasmodium vivax*. *BMC Genomics*, *14*(1), 8. https://doi.org/10.1186/1471-2164-14-8

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004075

Maccallum, W. G. (1897). On the flagellated form of the malarial parasite. *The Lancet*, *150*(3872), 1240–1241.

Madritsch, S., Burg, A., & Sehr, E. M. (2021). Comparing de novo transcriptome assembly tools in di- and autotetraploid non-model plant species. *BMC Bioinformatics*, *22*(1), 1–17. https://doi.org/10.1186/s12859-021-04078-8

Mair, G. R., Braks, J. A. M., Garver, L. S., Wiegant, J. C. A. G., Hall, N., Dirks, R. W., Khan, S. M., Dimopoulos, G., Janse, C. J., & Waters, A. P. (2006a). Regulation of sexual development of Plasmodium by translational repression. *Science*, *313*(5787), 667–669. https://doi.org/10.1126/science.1125129

Mair, G. R., Braks, J. A. M., Garver, L. S., Wiegant, J. C. A. G., Hall, N., Dirks, R. W., Khan, S. M., Dimopoulos, G., Janse, C. J., & Waters, A. P. (2006b). Translational Repression is essential for Plasmodium sexual development and mediated by a DDX6-type RNA helicase. *Science*, *313*(5787), 667. https://doi.org/10.1126/SCIENCE.1125129

Mancio-Silva, L., Gural, N., Real, E., Wadsworth, M. H., Butty, V. L., March, S., Nerurkar, N., Hughes, T. K., Roobsoong, W., Fleming, H. E., Whittaker, C. A., Levine, S. S., Sattabongkot, J., Shalek, A. K., & Bhatia, S. N. (2022). A single-cell liver atlas of *Plasmodium vivax* infection. *Cell Host and Microbe*, *30*(7), 1048-1060.e5. https://doi.org/10.1016/j.chom.2022.03.034

Mancio-Silva, L., Slavic, K., Grilo Ruivo, M. T., Grosso, A. R., Modrzynska, K. K., Vera, I. M., Sales-Dias, J., Gomes, A. R., Macpherson, C. R., Crozet, P., Adamo, M., Baena-Gonzalez, E., Tewari, R., Llinás, M., Billker, O., & Mota, M. M. (2017). Nutrient sensing modulates malaria parasite virulence. *Nature*, *547*(7662), 213–216. https://doi.org/10.1038/nature23009

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/MOLBEV/MSAB199

Marchiafava, E., & Bignami, A. (1892). *Sulle febbri malariche estivo-autunnali*. Innocenzo Artero.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: an open-source toolbox for large graph layout. *Visualization and Data Analysis 2011*, *7868*, 786806. https://doi.org/10.1117/12.871402

Matthews, H., McDonald, J., Totañes, F. I. G., & Merrick, C. J. (2022). Dynamics of DNA Replication during Male Gametogenesis in the Malaria Parasite Plasmodium Falciparum. *Cellular Microbiology*, *2022*, 1–17. https://doi.org/10.1155/2022/2701868

Mayne, B. (1932). Note on experimental infection of Anopheles punctipennis with quartan malaria. *Public Health Rep*, *47*, 1771–1773.

Ménard, D., Barnadas, C., Bouchier, C., Henry-Halldin, C., Gray, L. R., Ratsimbasoa, A., Thonier, V., Carod, J. F., Domarle, O., Colin, Y., Bertrand, O., Picot, J., King, C. L., Grimberg, B. T., Mercereau-Puijalon, O., & Zimmerman, P. A. (2010). Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(13), 5967–5971. https://doi.org/10.1073/PNAS.0912496107/SUPPL_FILE/PNAS.200912496SI.PDF

Meng, H., Yaari, G., Bolen, C. R., Avey, S., & Kleinstein, S. H. (2019). Gene set meta-analysis with Quantitative Set Analysis for Gene Expression (QuSAGE). *PLOS Computational Biology*, *15*(4), e1006899. https://doi.org/10.1371/journal.pcbi.1006899

Merino, E. F., Fernandez-Becerra, C., Durham, A. M., Ferreira, J. E., Tumilasci, V. F., d'Arc-Neves, J., da Silva-Nunes, M., Ferreira, M. U., Wickramarachchi, T., Udagama-Randeniya, P., Handunnetti, S. M., del Portillo, H. A., D'Arc-Neves, J., da Silva-Nunes, M., Ferreira, M. U., Wickramarachchi, T., Udagama-Randeniya, P., … del Portillo, H. A. (2006). Multi-character population study of the *vir* subtelomeric multigene superfamily of *Plasmodium vivax*, a major human malaria parasite. *Molecular and Biochemical Parasitology*, *149*(1), 10–16. https://doi.org/10.1016/j.molbiopara.2006.04.002

Miller, L. H., Mason, S. J., Clyde, D. F., & McGinniss, M. H. (1976). The Resistance Factor to *Plasmodium vivax* in Blacks. *New England Journal of Medicine*, *295*(6), 302–304. https://doi.org/10.1056/NEJM197608052950602

Minassian, A. M., Themistocleous, Y., Silk, S. E., Barrett, J. R., Kemp, A., Quinkert, D., Nielsen, C. M., Edwards, N. J., Rawlinson, T. A., Lopez, F. R., Roobsoong, W., Ellis, K. J. D., Cho, J. S., Aunin, E., Otto, T. D., Reid, A. J., Bach, F. A., … Draper, S. J. (2021). Controlled human malaria infection with a clone of *Plasmodium vivax* with high-quality genome assembly. *JCI Insight*, *6*(23). https://doi.org/10.1172/JCI.INSIGHT.152465

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/NAR/GKAA913

Modrzynska, K., Pfander, C., Chappell, L., Yu, L., Suarez, C., Dundas, K., Gomes, A. R., Goulding, D., Rayner, J. C., Choudhary, J., & Billker, O. (2017). A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the *Plasmodium* Life Cycle. *Cell Host and Microbe*, *21*(1), 11–22. https://doi.org/10.1016/j.chom.2016.12.003

Molina-Cruz, A., Garver, L. S., Alabaster, A., Bangiolo, L., Haile, A., Winikor, J., Ortega, C., van Schaijk, B. C. L., Sauerwein, R. W., Taylor-Salmon, E., & Barillas-Mury, C. (2013). The Human Malaria Parasite Pfs47 Gene Mediates Evasion of the Mosquito Immune System. *Science*, *340*(6135), 984–987. https://doi.org/10.1126/science.1235264

Muller, I., Jex, A. R., Kappe, S. H. I., Mikolajczak, S. A., Sattabongkot, J., Patrapuvich, R., Lindner, S., Flannery, E. L., Koepfli, C., Ansell, B., Lerch, A., Emery-Corbin, S. J., Charnaud, S., Smith, J., Merrienne, N., Swearingen, K. E., Moritz, R. L., … Chuenchob, V. (2019). Transcriptome and histone epigenome of *Plasmodium vivax* salivary-gland sporozoites point to tight regulatory control and mechanisms for liver-stage differentiation in relapsing malaria. *International Journal for Parasitology*, *49*(7), 501–513. https://doi.org/10.1016/j.ijpara.2019.02.007

Müller, K., & Wickham, H. (2019). *tibble: Simple Data Frames. R package version 2.1.3.*

Natalang, O., Bischoff, E., Deplaine, G., Proux, C., Dillies, M. A., Sismeiro, O., Guigon, G., Bonnefoy, S., Patarapotikul, J., Mercereau-Puijalon, O., Coppée, J. Y., & David, P. H. (2008). Dynamic RNA profiling in Plasmodium falciparum synchronized blood

stages exposed to lethal doses of artesunate. *BMC Genomics*, *9*, 388. https://doi.org/10.1186/1471-2164-9-388

Neafsey, D. E., Galinsky, K., Jiang, R. H. Y., Young, L., Sykes, S. M., Saif, S., Gujja, S., Goldberg, J. M., Young, S., Zeng, Q., Chapman, S. B., Dash, A. P., Anvikar, A. R., Sutton, P. L., Birren, B. W., Escalante, A. A., Barnwell, J. W., & Carlton, J. M. (2012). The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nature Genetics*, *44*(9), 1046–1050. https://doi.org/10.1038/ng.2373

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., … Besansky, N. J. (2015). Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, *347*(6217). https://doi.org/10.1126/SCIENCE.1258522

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, *94*(15), 7799–7806. https://doi.org/10.1073/pnas.94.15.7799

Nei, M., & Rooney, A. P. (2005). Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics*, *39*, 121. https://doi.org/10.1146/ANNUREV.GENET.39.073003.112240

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes. R package version*.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. https://doi.org/10.1038/nmeth.3337

Nurminsky, D. I., Nurminskaya, M. v., de Aguiar, D., & Hartl, D. L. (1998). Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature 1998 396:6711*, *396*(6711), 572–575. https://doi.org/10.1038/25126

Omelianczyk, R. I., Loh, H. P., Chew, M., Hoo, R., Baumgarten, S., Renia, L., Chen, J., & Preiser, P. R. (2020). Rapid activation of distinct members of multigene families in *Plasmodium spp*. *Communications Biology*, *3*(1), 351. https://doi.org/10.1038/s42003-020-1081-3

Otto, T. D., Böhme, U., Jackson, A. P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W. A. M., Religa, A. A., Robertson, L., Sanders, M., Ogun, S. A., Cunningham, D., Erhart, A., Billker, O., Khan, S. M., Stunnenberg, H. G., Langhorne, J., Holder, A. A., …

Janse, C. J. (2014). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biology*, *12*(1), 86. https://doi.org/10.1186/s12915-014-0086-0

Pain, A., Böhme, U., Berry, A. E., Mungall, K., Finn, R. D., Jackson, A. P., Mourier, T., Mistry, J., Pasini, E. M., Aslett, M. A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T. J., Cherevach, I., Chillingworth, T., … Berriman, M. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, *455*(7214), 799–803. https://doi.org/10.1038/nature07306

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pasini, E. M., Braks, J. A., Fonager, J., Klop, O., Aime, E., Spaccapelo, R., Otto, T. D., Berriman, M., Hiss, J. A., Thomas, A. W., Mann, M., Janse, C. J., Kocken, C. H. M., & Franke-Fayard, B. (2013). Proteomic and Genetic Analyses Demonstrate that *Plasmodium berghei* Blood Stages Export a Large and Diverse Repertoire of Proteins. *Molecular & Cellular Proteomics*, *12*(2), 426–448. https://doi.org/10.1074/mcp.M112.021238

Paton, M. G., Barker, G. C., Matsuoka, H., Ramesar, J., Janse, C. J., Waters, A. P., & Sinden, R. E. (1993). Structure and expression of a post-transcriptionally regulated malaria gene encoding a surface protein from the sexual stages of Plasmodium berghei. *Molecular and Biochemical Parasitology*, *59*(2), 263–275. https://doi.org/10.1016/0166-6851(93)90224-L

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*(359–367), 489–498. https://doi.org/10.1098/rspl.1896.0076

Pérez-Mazliah, D., Gardner, P. J., Schweighoffer, E., McLaughlin, S., Hosking, C., Tumwine, I., Davis, R. S., Potocnik, A. J., Tybulewicz, V. L., & Langhorne, J. (2018). Plasmodium-specific atypical memory B cells are short-lived activated B cells. *ELife*, *7*. https://doi.org/10.7554/eLife.39800

Petter, M., Bonow, I., & Klinkert, M. Q. (2008). Diverse Expression Patterns of Subgroups of the *rif* Multigene Family during *Plasmodium falciparum*

Gametocytogenesis. *PLoS ONE*, *3*(11). https://doi.org/10.1371/JOURNAL.PONE.0003779

Piehler, A. P., Hellum, M., Wenzel, J. J., Kaminski, E., Haug, K. B. F., Kierulf, P., & Kaminski, W. E. (2008). The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. *BMC Genomics*, *9*(1), 1–13. https://doi.org/10.1186/1471-2164-9-165/FIGURES/5

Poiani, A. (2006). Complexity of seminal fluid: A review. *Behavioral Ecology and Sociobiology*, *60*(3), 289–310. https://doi.org/10.1007/S00265-006-0178-0/FIGURES/1

Poinar, G. (2005). Plasmodium dominicana n. sp. (Plasmodiidae: Haemospororida) from Tertiary Dominican amber. *Systematic Parasitology*, *61*(1), 47–52. https://doi.org/10.1007/s11230-004-6354-6

Ponce, R., & Hartl, D. L. (2006). The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster*. *Gene*, *376*(2), 174–183. https://doi.org/10.1016/J.GENE.2006.02.011

Poudel, S. S., Newman, R. A., & Vaughan, J. A. (2008). Rodent Plasmodium: Population Dynamics of Early Sporogony within *Anopheles stephensi* Mosquitoes. *Journal of Parasitology*, *94*(5), 999–1008. https://doi.org/10.1645/GE-1407.1

Prieto-Godino, L. L., Rytz, R., Bargeton, B., Abuin, L., Arguello, J. R., Peraro, M. D., & Benton, R. (2016). Olfactory receptor pseudo-pseudogenes. *Nature*, *539*(7627), 93–97. https://doi.org/10.1038/nature19824

Prior, K. F., van der Veen, D. R., O'Donnell, A. J., Cumnock, K., Schneider, D., Pain, A., Subudhi, A., Ramaprasad, A., Rund, S. S. C., Savill, N. J., & Reece, S. E. (2018). Timing of host feeding drives rhythms in parasite replication. *PLOS Pathogens*, *14*(2), e1006900. https://doi.org/10.1371/journal.ppat.1006900

Qin, L., & Crawford, J. M. (2018). Anatomy and Cellular Functions of the Liver. *Zakim and Boyer's Hepatology*, 2-19.e4. https://doi.org/10.1016/B978-0-323-37591-7.00001-X

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/NAR/GKS1219

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Ramaprasad, A., Klaus, S., Douvropoulou, O., Culleton, R., & Pain, A. (2021). *Plasmodium vinckei* genomes provide insights into the pan-genome and evolution of rodent malaria parasites. *BMC Biology*, *19*(1), 69. https://doi.org/10.1186/s12915-021-00995-5

Ramaprasad, A., Subudhi, A. K., Culleton, R., & Pain, A. (2019). A fast and cost-effective microsampling protocol incorporating reduced animal usage for time-series transcriptomics in rodent malaria parasites. *Malaria Journal*, *18*(1), 26. https://doi.org/10.1186/s12936-019-2659-4

Rana, S. B., Zadlock, F. J., Zhang, Z., Murphy, W. R., & Bentivegna, C. S. (2016). Comparison of de Novo transcriptome assemblers and k-mer strategies using the killifish, fundulus heteroclitus. *PLoS ONE*, *11*(4), e0153104. https://doi.org/10.1371/journal.pone.0153104

Rangel, G. W., Clark, M. A., Kanjee, U., Goldberg, J. M., MacInnis, B., José Menezes, M., Ferreira, M. U., & Duraisingh, M. T. (2020). *Plasmodium vivax* transcriptional profiling of low input cryopreserved isolates through the intraerythrocytic development cycle. *PLOS Neglected Tropical Diseases*, *14*(3), e0008104. https://doi.org/10.1371/journal.pntd.0008104

Reid, A. J., Talman, A. M., Bennett, H. M., Gomes, A. R., Sanders, M. J., Illingworth, C. J. R., Billker, O., Berriman, M., & Lawniczak, M. K. N. (2018). Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *ELife*, *7*. https://doi.org/10.7554/eLife.33105

Reininger, L., Billker, O., Tewari, R., Mukhopadhyay, A., Fennell, C., Dorin-Semblat, D., Doerig, C., Goldring, D., Harmse, L., Ranford-Cartwright, L., Packer, J., & Doerig, C. (2005). A NIMA-related Protein Kinase Is Essential for Completion of the Sexual Cycle of Malaria Parasites. *Journal of Biological Chemistry*, *280*(36), 31957–31964. https://doi.org/10.1074/JBC.M504523200

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Risal, P., Cho, B. H., Sylvester, K. G., Kim, J.-C., Kim, H. T., & Jeong, Y. J. (2011). The establishment and characterization of immortal hepatocyte cell lines from a mouse liver injury model. *In Vitro Cellular & Developmental Biology - Animal 2011 47:8*, *47*(8), 526–534. https://doi.org/10.1007/S11626-011-9445-0

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, *32*(9), 896–902. https://doi.org/10.1038/nbt.2931

Rivera-Vicéns, R. E., Garcia-Escudero, C. A., Conci, N., Eitel, M., & Wörheide, G. (2022). TransPi—a comprehensive TRanscriptome ANalysiS PIpeline for de novo transcriptome assembly. *Molecular Ecology Resources*, *22*(5), 2070–2086. https://doi.org/10.1111/1755-0998.13593

Rivero, F. D., Saura, A., Prucca, C. G., Carranza, P. G., Torri, A., & Lujan, H. D. (2010). Disruption of antigenic variation is crucial for effective parasite vaccine. *Nature Medicine 2010 16:5*, *16*(5), 551–557. https://doi.org/10.1038/nm.2141

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods 2010 7:11*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

Robertson, S. A., Guerin, L. R., Bromfield, J. J., Branson, K. M., Ahlström, A. C., & Care, A. S. (2009). Seminal fluid drives expansion of the CD4+CD25+ T regulatory cell pool and induces tolerance to paternal alloantigens in mice. *Biology of Reproduction*, *80*(5), 1036–1045. https://doi.org/10.1095/BIOLREPROD.108.074658

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Roch, K. G. Le, 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*(11), 2308. https://doi.org/10.1101/GR.2523904

Rorick, M. M., Rask, T. S., Baskerville, E. B., Day, K. P., & Pascual, M. (2013). Homology blocks of Plasmodium falciparum var genes and clinically distinct forms of severe malaria in a local population. *BMC Microbiology*, *13*(1), 1. https://doi.org/10.1186/1471-2180-13-244

Ross, R. (1897). On some Peculiar Pigmented Cells Found in Two Mosquitos Fed on Malarial Blood. *British Medical Journal*, *2*(1929), 1786. https://doi.org/10.1136/BMJ.2.1929.1786

Ross, R. (1898). THE ROLE OF THE MOSQUITO IN THE EVOLUTION OF THE MALARIAL PARASITE. *The Lancet*, *152*(3912), 488–490. https://doi.org/10.1016/S0140-6736(01)81400-8

Roth, A., Adapa, S. R., Zhang, M., Liao, X., Saxena, V., Goffe, R., Li, S., Ubalee, R., Saggu, G. S., Pala, Z. R., Garg, S., Davidson, S., Jiang, R. H. Y., & Adams, J. H. (2018). Unraveling the *Plasmodium vivax* sporozoite transcriptional journey from mosquito vector to human host. *Scientific Reports*, *8*(1), 12183. https://doi.org/10.1038/s41598-018-30713-1

Rutledge, G. G., Böhme, U., Sanders, M., Reid, A. J., Cotton, J. A., Maiga-Ascofare, O., Djimdé, A. A., Apinjoh, T. O., Amenga-Etego, L., Manske, M., Barnwell, J. W., Renaud, F., Ollomo, B., Prugnolle, F., Anstey, N. M., Auburn, S., Price, R. N., … Otto, T. D. (2017). *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature*, *542*(7639), 101–104. https://doi.org/10.1038/nature21038

Sà, J. M., Cannon, M. V., Caleon, R. L., Wellems, T. E., & Serre, D. (2020). Single-cell transcription analysis of *Plasmodium vivax* blood-stage parasites identifies stage- And species-specific profiles of expression. *PLoS Biology*, *18*(5), e3000711. https://doi.org/10.1371/journal.pbio.3000711

Saito, F., Hirayasu, K., Satoh, T., Wang, C. W., Lusingu, J., Arimori, T., Shida, K., Palacpac, N. M. Q., Itagaki, S., Iwanaga, S., Takashima, E., Tsuboi, T., Kohyama, M., Suenaga, T., Colonna, M., Takagi, J., Lavstsen, T., … Arase, H. (2017). Immune evasion of Plasmodium falciparum by RIFIN via inhibitory receptors. *Nature*, 1–49. https://doi.org/10.1038/nature24994

Sanni, L. A., Jarra, W., Li, C., & Langhorne, J. (2004). Cerebral Edema and Cerebral Hemorrhages in Interleukin-10-Deficient Mice Infected with Plasmodium chabaudi. *Infection and Immunity*, *72*(5), 3054–3058. https://doi.org/10.1128/IAI.72.5.3054-3058.2004/ASSET/F0094108-0747-484D-99D5-FCEAA0FCBAA6/ASSETS/GRAPHIC/ZII0050439140004.JPEG

Sato, Y., Montagna, G. N., & Matuschewski, K. (2014). Plasmodium berghei sporozoites acquire virulence and immunogenicity during mosquito hemocoel transit. *Infection and Immunity*, *82*(3), 1164–1172. https://doi.org/10.1128/IAI.00758-13

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864. https://doi.org/10.1101/GR.213611.116

Shaw, P. J., Chaotheing, S., Kaewprommal, P., Piriyapongsa, J., Wongsombat, C., Suwannakitti, N., Koonyosying, P., Uthaipibull, C., Yuthavong, Y., & Kamchonwongpaisan, S. (2015). *Plasmodium* parasites mount an arrest response to dihydroartemisinin, as revealed by whole transcriptome shotgun sequencing (RNA-seq) and microarray study. *BMC Genomics*, *16*(1). https://doi.org/10.1186/S12864-015-2040-0

Shears, M. J., Sekhar Nirujogi, R., Swearingen, K. E., Renuse, S., Mishra, S., Jaipal Reddy, P., Moritz, R. L., Pandey, A., & Sinnis, P. (2019). Proteomic Analysis of *Plasmodium* Merosomes: The Link between Liver and Blood Stages in Malaria. *Journal of Proteome Research*, acs.jproteome.9b00324. https://doi.org/10.1021/acs.jproteome.9b00324

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/JOURNAL.PONE.0163962

Shortt, H. E., & Garnham, P. C. C. (1948). Pre-erythrocytic Stage in Mammalian Malaria Parasites. *Nature*, *161*(4082), 126–126. https://doi.org/10.1038/161126a0

Shortt, H. E., Garnham, P. C. C., Covell, G., & Shute, P. G. (1948). Pre-Erythrocytic Stage of Human Malaria, *Plasmodium Vivax*. *Br Med J*, *1*(4550), 547–547. https://doi.org/10.1136/BMJ.1.4550.547

Sieber, K. P., Huber, M., Kaslow, D., Banks, S. M., Torii, M., Aikawa, M., & Miller, L. H. (1991). The peritrophic membrane as a barrier: Its penetration by *Plasmodium gallinaceum* and the effect of a monoclonal antibody to ookinetes. *Experimental Parasitology*, *72*(2), 145–156. https://doi.org/10.1016/0014-4894(91)90132-G

Siegel, S. V., Chappell, L., Hostetler, J. B., Amaratunga, C., Suon, S., Böhme, U., Berriman, M., Fairhurst, R. M., & Rayner, J. C. (2020). Analysis of *Plasmodium vivax* schizont transcriptomes from field isolates reveals heterogeneity of expression of genes involved in host-parasite interactions. *Scientific Reports*, *10*(1), 16667. https://doi.org/10.1038/s41598-020-73562-7

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com

Silveira, H., Ramos, S., Abrantes, P., Lopes, L. F., do Rosario, V. E., & Abrahamsen, M. S. (2007). Effect of chloroquine on gene expression of Plasmodium yoelii nigeriensis during its sporogonic development in the mosquito vector. *Malaria Journal*, *6*, 84. https://doi.org/10.1186/1475-2875-6-84

Silvie, O., Briquet, S., Müller, K., Manzoni, G., & Matuschewski, K. (2014). Post-transcriptional silencing of UIS4 in Plasmodium berghei sporozoites is important for host switch. *Molecular Microbiology*, *91*(6), 1200–1213. https://doi.org/10.1111/MMI.12528

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. v., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sinden, R. E., Butcher, G. A., & Beetsma, A. L. (2002). Maintenance of the Plasmodium berghei life cycle. *Methods in Molecular Medicine*, *72*, 25–40. https://doi.org/10.1385/1-59259-271-6:25

Sinden, R. E., Talman, A., Marques, S. R., Wass, M. N., & Sternberg, M. J. E. (2010). The flagellum in malarial parasites. *Current Opinion in Microbiology*, *13*(4), 491–500. https://doi.org/10.1016/J.MIB.2010.05.016

Smith, J. D., Chitnis, C. E., Craig, A. G., Roberts, D. J., Hudson-Taylor, D. E., Peterson, D. S., Pinches, R., Newbold, C. I., & Miller, L. H. (1995). Switches in Expression of *Plasmodium falciparum var* Genes Correlate with Changes in Antigenic and Cytoadherent Phenotypes of Infected Erythrocytes. *Cell*, *82*(1), 101. https://doi.org/10.1016/0092-8674(95)90056-X

Smith, J. D., Subramanian, G., Gamain, B., Baruch, D. I., & Miller, L. H. (2000). Classification of adhesive domains in the Plasmodium falciparum Erythrocyte Membrane Protein 1 family. *Molecular and Biochemical Parasitology*, *110*(2), 293–310. https://doi.org/10.1016/S0166-6851(00)00279-6

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8), 1134. https://doi.org/10.1101/GR.196469.115

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray

hybridization. *Molecular Biology of the Cell*, *9*(12), 3273–3297. https://doi.org/10.1091/MBC.9.12.3273

Spence, P. J., Jarra, W., Lévy, P., Nahrendorf, W., & Langhorne, J. (2012). Mosquito transmission of the rodent malaria parasite *Plasmodium chabaudi*. *Malaria Journal*, *11*(1), 407. https://doi.org/10.1186/1475-2875-11-407

Spence, P. J., Jarra, W., Lévy, P., Reid, A. J., Chappell, L., Brugat, T., Sanders, M., Berriman, M., & Langhorne, J. (2013). Vector transmission regulates immune control of *Plasmodium* virulence. *Nature*, *498*(7453), 228–231. https://doi.org/10.1038/nature12231

SRA_Toolkit_Development_Team. (n.d.). *SRA-Tools*. https://github.com/ncbi/sra-tools

Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics*, *18*(2), kxw041. https://doi.org/10.1093/biostatistics/kxw041

Stephens, R., Culleton, R. L., & Lamb, T. J. (2012). The contribution of *Plasmodium chabaudi* to our understanding of malaria. *Trends in Parasitology*, *28*(2), 73–82. https://doi.org/10.1016/j.pt.2011.10.006

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*(5791), 1287–1290. https://doi.org/10.1126/science.1129720

Su, X., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., Ravetch, J. A., & Wellems, T. E. (1995). The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell*, *82*(1), 89–100. https://doi.org/10.1016/0092-8674(95)90055-1

Subudhi, A. K., O'Donnell, A. J., Ramaprasad, A., Abkallo, H. M., Kaushik, A., Ansari, H. R., Abdel-Haleem, A. M., ben Rached, F., Kaneko, O., Culleton, R., Reece, S. E., & Pain, A. (2020). Malaria parasites regulate intra-erythrocytic development duration via serpentine receptor 10 to coordinate with host rhythms. *Nature Communications*, *11*(1), 1–15. https://doi.org/10.1038/s41467-020-16593-y

Swanson, W. J., & Vacquier, V. D. (2002). The rapid evolution of reproductive proteins. *Nature Reviews Genetics 2002 3:2*, *3*(2), 137–144. https://doi.org/10.1038/nrg733

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Thurston, J. P. (1951). Morphological changes in Plasmodium berghei following proguanil, sulphadiazine and mepacrine therapy. *Transactions of the Royal Society*

of Tropical Medicine and Hygiene, *44*(6), 703–706. https://doi.org/10.1016/0035-9203(51)90006-5

Tusting, L. S., Willey, B., Lucas, H., Thompson, J., Kafy, H. T., Smith, R., & Lindsay, S. W. (2013). Socioeconomic development as an intervention against malaria: a systematic review and meta-analysis. *The Lancet*, *382*(9896), 963–972. https://doi.org/10.1016/S0140-6736(13)60851-X

Ukegbu, C. V., Giorgalli, M., Yassine, H., Ramirez, J. L., Taxiarchi, C., Barillas-Mury, C., Christophides, G. K., & Vlachou, D. (2017). Plasmodium berghei P47 is essential for ookinete protection from the Anopheles gambiae complement-like response. *Scientific Reports*, *7*(1), 6026. https://doi.org/10.1038/s41598-017-05917-6

van den Berghe, L., Vincke, I., & Chardome, M. (1950). *La phase tissulaire de Plasmodium berghei*. *Annales de la Societe belge de medecine tropicale (1920)*, *30*(1), 79–82. http://www.ncbi.nlm.nih.gov/pubmed/15413859

van Dijk, M. R., van Schaijk, B. C. L., Khan, S. M., van Dooren, M. W., Ramesar, J., Kaczanowski, S., van Gemert, G. J. van, Kroeze, H., Stunnenberg, H. G., Eling, W. M., Sauerwein, R. W., Waters, A. P., & Janse, C. J. (2010). Three Members of the *6-cys* Protein Family of *Plasmodium* Play a Role in Gamete Fertility. *PLOS Pathogens*, *6*(4), e1000853. https://doi.org/10.1371/JOURNAL.PPAT.1000853

van Dongen, S., & Abreu-Goodger, C. (2012). Using MCL to Extract Clusters from Networks. *Methods in Molecular Biology*, *804*, 281–295. https://doi.org/10.1007/978-1-61779-361-5_15

Vaughan, J. A. (2007). Population dynamics of *Plasmodium* sporogony. *Trends in Parasitology*, *23*(2), 63–70. https://doi.org/10.1016/j.pt.2006.12.009

Vincke, I. H., & Lips, M. (1948). *Un nouveau plasmodium d'un rongeur sauvage du Congo, Plasmodium berghei n. sp. Annales de La Société Belge de Médecine Tropicale*, *Mar 28*(1), 97–104.

von der Weid, T., Honarvar, N., & Langhorne, J. (1996). Gene-targeted mice lacking B cells are unable to eliminate a blood stage malaria infection. *The Journal of Immunology*, *156*(7).

Voorberg-van der Wel, A., Roma, G., Gupta, D. K., Schuierer, S., Nigsch, F., Carbone, W., Zeeman, A.-M., Lee, B. H., Hofman, S. O., Faber, B. W., Knehr, J., Pasini, E., Kinzel, B., Bifani, P., Bonamy, G. M. C. C., Bouwmeester, T., Kocken, C. H. M. M., & Diagana, T. T. (2017). A comparative transcriptomic analysis of replicating and dormant liver stages of the relapsing malaria parasite *Plasmodium cynomolgi*. *ELife*, *6*(1). https://doi.org/10.7554/eLife.29605

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, *131*(4), 281–285. https://doi.org/10.1007/s12064-012-0162-3

Wang, L. G., Lam, T. T. Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., & Yu, G. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, *37*(2), 599–603. https://doi.org/10.1093/molbev/msz240

Wei, T., & Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix (Version 0.84)*. https://github.com/taiyun/corrplot

Wéry, M. (1967). *Etude du cycle sporogonique de* Plasmodium chabaudi *en vue de la production massive de sporozoïtes viables et de formes exo-érythrocytaires*. *Protistologica*, *3*, 275–285.

WHO. (2016). *Malaria Free Sri Lanka*.

WHO. (2021). *World Malaria Report 2021*.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag.

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0.* https://cran.r-project.org/package=stringr

Wickham, H., & Bryan, J. (2019). *readxl: Read Excel Files. R package* (1.3.1). https://cran.r-project.org/package=readxl

Wickham, H., Hester, J., & Francois, R. (2018). *readr: Read Rectangular Text Data. R package* (1.3.1). https://cran.r-project.org/package=readr

Wickham, H., Romain, F., Henry, L., & Kirill, M. (2019). *dplyr: A Grammar of Data Manipulation. R package version 0.8.1.* https://cran.r-project.org/package=dplyr

Witmer, K., Fraschka, S. A., Vlachou, D., Bártfai, R., & Christophides, G. K. (2020). An epigenetic map of malaria parasite development from host to vector. *Scientific Reports*, *10*(1), 1–19. https://doi.org/10.1038/s41598-020-63121-5

Wright, J. C., Mudge, J., Weisser, H., Barzine, M. P., Gonzalez, J. M., Brazma, A., Choudhary, J. S., & Harrow, J. (2016). Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature Communications*, *7*. https://doi.org/10.1038/NCOMMS11778

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G. K.-S., & Wang, J. (2014).

SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. https://doi.org/10.1093/BIOINFORMATICS/BTU077

Yahav, T., & Privman, E. (2019). A comparative analysis of methods for de novo assembly of hymenopteran genomes using either haploid or diploid samples. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-42795-6

Yam, X. Y., Brugat, T., Siau, A., Lawton, J., Wong, D. S., Farah, A., Twang, J. S., Gao, X., Langhorne, J., & Preiser, P. R. (2016). Characterization of the Plasmodium Interspersed Repeats (PIR) proteins of *Plasmodium chabaudi* indicates functional diversity. *Scientific Reports*, *6*(March), 23449. https://doi.org/10.1038/srep23449

Yeh, S. D., Do, T., Chan, C., Cordova, A., Carranza, F., Yamamoto, E. A., Abbassi, M., Gandasetiawan, K. A., Librado, P., Damia, E., Dimitri, P., Rozas, J., Hartl, D. L., Roote, J., & Ranz, J. M. (2012). Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), 2043–2048. https://doi.org/10.1073/PNAS.1121327109

Yin, D., Schwarz, E. M., Thomas, C. G., Felde, R. L., Korf, I. F., Cutter, A. D., Schartner, C. M., Ralston, E. J., Meyer, B. J., & Haag, E. S. (2018). Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science*, *359*(6371), 55–61. https://doi.org/10.1126/SCIENCE.AAO0827

Yoeli, M., & Hargreaves, B. J. (1974). Brain capillary blockage produced by a virulent strain of rodent malaria. *Science*, *184*(4136), 572–573.

Yoeli, M., Nussenzweig, R., Upmanis, R. S., & Most, H. (1966). Resistance of *Plasmodium chabaudi*-Infected White Mice to a Fulminating and Fatal Strain of *Plasmodium vinckei*. *Nature*, *211*(5044), 49–51. https://doi.org/10.1038/211049a0

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Yuda, M., Iwanaga, S., Kaneko, I., & Kato, T. (2015). Global transcriptional repression: An initial and essential step for *Plasmodium* sexual development. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12824–12829. https://doi.org/10.1073/pnas.1504389112

Zanghì, G., Vembar, S. S., Baumgarten, S., Ding, S., Guizetti, J., Bryant, J. M., Mattei, D., Jensen, A. T. R., Rénia, L., Goh, Y. S., Sauerwein, R., Hermsen, C. C.,

Franetich, J. F., Bordessoulles, M., Silvie, O., Soulard, V., Scatton, O., … Scherf, A. (2018). A Specific PfEMP1 Is Expressed in *P. falciparum* Sporozoites and Plays a Role in Hepatocyte Infection. *Cell Reports*. https://doi.org/10.1016/j.celrep.2018.02.075

Zhang, C., Oguz, C., Huse, S., Xia, L., Wu, J., Peng, Y. C., Smith, M., Chen, J., Long, C. A., Lack, J., & Su, X. zhuan. (2021). Genome sequence, transcriptome, and annotation of rodent malaria parasite *Plasmodium yoelii nigeriensis* N67. *BMC Genomics*, *22*(1), 1–12. https://doi.org/10.1186/s12864-021-07555-9

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, *2*(3). https://doi.org/10.1093/NARGAB/LQAA078

Zhu, L., Mok, S., Imwong, M., Jaidee, A., Russell, B., Nosten, F., Day, N. P., White, N. J., Preiser, P. R., & Bozdech, Z. (2016). New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. *Scientific Reports*, *6*(1), 20498. https://doi.org/10.1038/srep20498

Zhu, L., van der Pluijm, R. W., Kucharski, M., Nayak, S., Tripathi, J., White, N. J., Day, N. P. J., Faiz, A., Phyo, A. P., Amaratunga, C., Lek, D., Ashley, E. A., Nosten, F., Smithuis, F., Ginsburg, H., von Seidlein, L., Lin, K., … Bozdech, Z. (2022). Artemisinin resistance in the malaria parasite, *Plasmodium falciparum*, originates from its initial transcriptional response. *Communications Biology 2022 5:1*, *5*(1), 1–13. https://doi.org/10.1038/s42003-022-03215-0

Chapter 8. Appendix

Experimental Code	Samples names	Cat Sample stage(s)	Concatenated Files	Rationale	Geographic source	Run accession
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571697
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571698
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571699
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571700
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571701
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571702
E01	SMRU1	Asexual_Mixed	SMRU1	Clinical isolate SMRU1	NW Thailand _ Mae Sot	SRR1571703
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571704
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571705
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571706
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571707
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571708
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571709
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571710
E01	SMRU2	Asexual_Mixed	SMRU2	Clinical isolate SMRU2	NW Thailand _ Mae Sot	SRR1571711
E01	Mixed	Asexual_Mixed	Mixed	Mixed SMRU1 and SMRU2	NW Thailand _ Mae Sot	SRR1571712
E01	Mixed	Asexual_Mixed	Mixed	Mixed SMRU1 and SMRU2	NW Thailand _ Mae Sot	SRR1571713
E01	Mixed	Asexual_Mixed	Mixed	Mixed SMRU1 and SMRU2	NW Thailand _ Mae Sot	SRR1571714
E01	Mixed	Asexual_Mixed	Mixed	Mixed SMRU1 and SMRU2	NW Thailand _ Mae Sot	SRR1571715
E02	V_DJK_8	Asexual_Mixed		Separate clinical isolates	Cambodia	SRR5646671
E02	V_DJK_10	Asexual_Mixed		Separate clinical isolates	Cambodia	SRR5646670

E02	V_DJK_16	Asexual_Mixed		Separate clinical isolates	Cambodia	SRR5646669
E02	ER_0 before CQ (no after)	Asexual_Mixed		Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347533
E02	AR_0 before CQ (no after)	Asexual_Mixed		Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347532
E02	FR_0 before CQ (no after)	Asexual_Mixed		Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347510
E02	Sp_1 Anopheles albimanus	Sporozoite		Separate clinical isolates	Cambodia	SRR5646668
E03	Hyp1	Liver_Hypnozoite		*Difficult to tell?*	Thailand	SRR6371896
E03	Hyp2	Liver_Hypnozoite		*Difficult to tell?*	Thailand	SRR6371897
E03	Mixed1	Liver_Mixed		*Difficult to tell?*	Thailand	SRR6371894
E03	Mixed2	Liver_Mixed		*Difficult to tell?*	Thailand	SRR6371895
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand a	ERR2541707
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand a	ERR2541708
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand a	ERR2541709
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand a	ERR2541710
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand a	ERR2541711
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand b	ERR2541712
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand b	ERR2541713
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand b	ERR2541714
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand b	ERR2541715
E04	None	Asexual_Mixed		Separate clinical isolates	N or S Thailand b	ERR2541716
E05	Thai1	Sporozoite		Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298190

E05	Thai2	Sporozoite	Thai2	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298184
E05	Thai2	Sporozoite	Thai2	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298185
E05	Thai3	Sporozoite	Thai3	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298182
E05	Thai3	Sporozoite	Thai3	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298183
E05	Thai4	Sporozoite	Thai4	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298180
E05	Thai4	Sporozoite	Thai4	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298181
E05	Thai5	Sporozoite	Thai5	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298178
E05	Thai5	Sporozoite	Thai5	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298179
E05	Thai6	Sporozoite	Thai6	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298188
E05	Thai6	Sporozoite	Thai6	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298189
E05	Thai7	Sporozoite	Thai7	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298176
E05	Thai7	Sporozoite	Thai7	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298177
E05	Thai8	Sporozoite	Thai8	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298174
E05	Thai8	Sporozoite	Thai8	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298175
E05	Thai9	Sporozoite	Thai9	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298172
E05	Thai9	Sporozoite	Thai9	Separate clinical isolates	Thailand _ Tak/Ubon	SRR5298173
E06	A_0 before CQ	Asexual_Mixed	Kim19_A	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347515
E06	A_1 after CQ	Asexual_Mixed	Kim19_A	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347514
E06	J_0	Asexual_Mixed	Kim19_J	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347502
E06	J_1	Asexual_Mixed	Kim19_J	Separate clinical	Cambodia	SRR8347501

				isolates - before and after CQ for each patient combined		
E06	K_0	Asexual_Mixed	Kim19_K	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347526
E06	K_1	Asexual_Mixed	Kim19_K	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347527
E06	L_0	Asexual_Mixed	Kim19_L	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347528
E06	L_1	Asexual_Mixed	Kim19_L	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347529
E06	M_0	Asexual_Mixed	Kim19_M	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347522
E06	M_1	Asexual_Mixed	Kim19_M	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347523
E06	N_0	Asexual_Mixed	Kim19_N	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347524
E06	N_1	Asexual_Mixed	Kim19_N	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347525
E06	O_0	Asexual_Mixed	Kim19_O	Separate clinical isolates - before and	Cambodia	SRR8347530

				after CQ for each patient combined		
E06	O_1	Asexual_Mixed	Kim19_O	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347531
E06	P_0	Asexual_Mixed	Kim19_P	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347504
E06	P_1	Asexual_Mixed	Kim19_P	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347503
E06	Q_0	Asexual_Mixed	Kim19_Q	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347518
E06	Q_1	Asexual_Mixed	Kim19_Q	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347509
E06	S_0	Asexual_Mixed	Kim19_S	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347508
E06	S_1	Asexual_Mixed	Kim19_S	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347507
E06	T_0	Asexual_Mixed	Kim19_T	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347506
E06	T_1	Asexual_Mixed	Kim19_T	Separate clinical isolates - before and after CQ for each	Cambodia	SRR8347505

				patient combined		
E06	B_0	Asexual_Mixed	Kim19_B	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347513
E06	B_1	Asexual_Mixed	Kim19_B	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347512
E06	U_0	Asexual_Mixed	Kim19_U	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347535
E06	U_1	Asexual_Mixed	Kim19_U	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347534
E06	C_0	Asexual_Mixed	Kim19_C	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347519
E06	C_1	Asexual_Mixed	Kim19_C	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347511
E06	D_0	Asexual_Mixed	Kim19_D	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347517
E06	D_1	Asexual_Mixed	Kim19_D	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347516
E06	E_0	Asexual_Mixed	Kim19_E	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347521

E06	E_1	Asexual_Mixed	Kim19_E	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347520
E06	F_0	Asexual_Mixed	Kim19_F	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347494
E06	F_1	Asexual_Mixed	Kim19_F	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347493
E06	G_0	Asexual_Mixed	Kim19_G	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347496
E06	G_1	Asexual_Mixed	Kim19_G	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347495
E06	H_0	Asexual_Mixed	Kim19_H	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347498
E06	H_1	Asexual_Mixed	Kim19_H	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347497
E06	I_0	Asexual_Mixed	Kim19_I	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347500
E06	I_1	Asexual_Mixed	Kim19_I	Separate clinical isolates - before and after CQ for each patient combined	Cambodia	SRR8347499
E07		Asexual_Mixed	PRJNA515743_Guna19	Same strain	Salvador (Sal1)	SRR8452526

E07		Asexual_Mixed	PRJNA515743 _Guna19	Same strain	Salvador (Sal1)	SRR8452527
E07		Asexual_Mixed	PRJNA515743 _Guna19	Same strain	Salvador (Sal1)	SRR8452528
E07		Asexual_Mixed	PRJNA515743 _Guna19	Same strain	Salvador (Sal1)	SRR8452529
E08	Infected Songkhla 1 Day 7	Oocyst		Two separate locations	Thailand _ Songkhla	SRR10106503
E08	Infected Yala Day 7	Oocyst		Two separate locations	Thailand _ Yala	SRR10106504
E08	Infected_2 18h Songkhla	Ookinete		Two separate locations	Thailand _ Songkhla	SRR10149834
E08	Infected_1 18h Yala	Ookinete		Two separate locations	Thailand _ Yala	SRR10149833
E09	None	Asexual_Schizont		Separate clinical isolates	Cambodia	ERR1717084
E09	None	Asexual_Schizont		Separate clinical isolates	Cambodia	ERR1717085
E09	None	Asexual_Schizont		Separate clinical isolates	Cambodia	ERR1717086
E09	None	Asexual_Schizont		Separate clinical isolates	Cambodia	ERR1717087
E10	PvJBC_Aim V_20_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203168
E10	PvJBC_Aim V_20_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203284
E10	PvJBC_Aim V_36_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203176
E10	PvJBC_Aim V_36_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203158
E10	PvJBC_Aim V_4_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203194
E10	PvJBC_Aim V_4_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203273
E10	PvJBC_Aim V_4_3	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203254
E10	PvJBC_DMEM_20_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203207
E10	PvJBC_DMEM_20_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203187
E10	PvJBC_DMEM_36_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203272
E10	PvJBC_DMEM_36_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203253
E10	PvJBC_DMEM_4_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203231
E10	PvJBC_DMEM_4_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203230
E10	PvJBC_DMEM_4_3	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203214
E10	PvJBC_IMDM_20_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203266
E10	PvJBC_IMDM_20_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203248
E10	PvJBC_IMDM_36_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203215
E10	PvJBC_IMDM_36_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203195
E10	PvJBC_IMDM_4_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203196

E10	PvJBC_IMDM_4_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203177
E10	PvJBC_IMDM_4_3	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203159
E10	PvJBC_McCoy's 5A_20_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203171
E10	PvJBC_McCoy's 5A_20_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203152
E10	PvJBC_McCoy's 5A_36_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203280
E10	PvJBC_McCoy's 5A_36_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203232
E10	PvJBC_McCoy's 5A_4_1	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203281
E10	PvJBC_McCoy's 5A_4_2	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203233
E10	PvJBC_McCoy's 5A_4_3	Asexual_Mixed	PvJBC	Three bio replicates	Brazil _ Acre	SRR10203216
E10	PvMRMS_Aim V_20_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203169
E10	PvMRMS_Aim V_20_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203150
E10	PvMRMS_Aim V_36_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203223
E10	PvMRMS_Aim V_36_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203229
E10	PvMRMS_Aim V_4_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203175
E10	PvMRMS_Aim V_4_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203156
E10	PvMRMS_Aim V_4_3	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203271
E10	PvMRMS_Aim V_44 DNA high_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203179
E10	PvMRMS_Aim V_44 DNA high_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203225
E10	PvMRMS_Aim V_44 DNA low_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203164
E10	PvMRMS_Aim V_44 DNA low_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203278
E10	PvMRMS_DMEM_20_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203264
E10	PvMRMS_DMEM_20_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203245
E10	PvMRMS_DMEM_36_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203213

E10	PvMRMS_DMEM_36_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203193
E10	PvMRMS_DMEM_4_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203212
E10	PvMRMS_DMEM_4_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203252
E10	PvMRMS_IMDM_20_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203205
E10	PvMRMS_IMDM_20_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203185
E10	PvMRMS_IMDM_36_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203270
E10	PvMRMS_IMDM_36_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203251
E10	PvMRMS_IMDM_4_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203192
E10	PvMRMS_IMDM_44 DNA high_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203170
E10	PvMRMS_IMDM_44 DNA high_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203151
E10	PvMRMS_IMDM_44 DNA high_3	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203265
E10	PvMRMS_IMDM_44 DNA low_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203237
E10	PvMRMS_IMDM_44 DNA low_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203220
E10	PvMRMS_IMDM_44 DNA low_3	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203200
E10	PvMRMS_McCoy's 5A_20_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203279
E10	PvMRMS_McCoy's 5A_20_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203282
E10	PvMRMS_McCoy's 5A_36_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203174
E10	PvMRMS_McCoy's 5A_36_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203155
E10	PvMRMS_McCoy's 5A_4_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203228
E10	PvMRMS_McCoy's 5A_4_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203211
E10	PvMRMS_McCoy's 5A_44 DNA high_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203208
E10	PvMRMS_McCoy's	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203188

	5A_44 DNA high_2					
E10	PvMRMS_ McCoy's 5A_44 DNA low_1	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203260
E10	PvMRMS_ McCoy's 5A_44 DNA low_2	Asexual_Mixed	PvMRMS	Three bio replicates	Brazil _ Acre	SRR10203241
E10	PvTCF_Aim V_20_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203203
E10	PvTCF_Aim V_20_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203183
E10	PvTCF_Aim V_36_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203267
E10	PvTCF_Aim V_36_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203249
E10	PvTCF_Aim V_4_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203209
E10	PvTCF_Aim V_4_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203189
E10	PvTCF_Aim V_44 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203268
E10	PvTCF_Aim V_44 DNA high_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203263
E10	PvTCF_Aim V_44 DNA high_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203244
E10	PvTCF_Aim V_44 DNA low_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203161
E10	PvTCF_Aim V_44 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203275
E10	PvTCF_Aim V_44 DNA low_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203256
E10	PvTCF_Aim V_72 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203219
E10	PvTCF_Aim V_72 DNA high_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203199
E10	PvTCF_Aim V_72 DNA mid_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203236
E10	PvTCF_Aim V_72 DNA mid_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203240
E10	PvTCF_DMEM_20_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203257
E10	PvTCF_DMEM_20_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203222

E10	PvTCF_DMEM_36_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203172
E10	PvTCF_DMEM_36_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203153
E10	PvTCF_DMEM_4_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203190
E10	PvTCF_DMEM_4_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203226
E10	PvTCF_DMEM_44 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203149
E10	PvTCF_DMEM_44 DNA low_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203218
E10	PvTCF_DMEM_44 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203198
E10	PvTCF_DMEM_44 DNA low_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203180
E10	PvTCF_DMEM_72 DNA low_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203160
E10	PvTCF_DMEM_72 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203274
E10	PvTCF_DMEM_72 DNA mid_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203163
E10	PvTCF_DMEM_72 DNA mid_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203277
E10	PvTCF_DMEM_72 DNA mid_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203259
E10	PvTCF_IMDM_20_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203166
E10	PvTCF_IMDM_20_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203148
E10	PvTCF_IMDM_36_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203201
E10	PvTCF_IMDM_36_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203227
E10	PvTCF_IMDM_4_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203173
E10	PvTCF_IMDM_4_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203154
E10	PvTCF_IMDM_44 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203157
E10	PvTCF_IMDM_44 DNA high_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203283
E10	PvTCF_IMDM_44 DNA high_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203247
E10	PvTCF_IMDM_44	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203181

	DNA low_1					
E10	PvTCF_IMDM_44 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203162
E10	PvTCF_IMDM_44 DNA low_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203276
E10	PvTCF_IMDM_72 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203184
E10	PvTCF_IMDM_72 DNA high_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203165
E10	PvTCF_IMDM_72 DNA high_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203147
E10	PvTCF_IMDM_72 DNA low_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203234
E10	PvTCF_IMDM_72 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203238
E10	PvTCF_IMDM_72 DNA mid_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203255
E10	PvTCF_IMDM_72 DNA mid_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203224
E10	PvTCF_IMDM_72 DNA mid_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203204
E10	PvTCF_McCoy's 5A_20_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203262
E10	PvTCF_McCoy's 5A_20_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203243
E10	PvTCF_McCoy's 5A_36_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203210
E10	PvTCF_McCoy's 5A_36_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203191
E10	PvTCF_McCoy's 5A_4_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203269
E10	PvTCF_McCoy's 5A_4_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203250
E10	PvTCF_McCoy's 5A_44 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203206
E10	PvTCF_McCoy's 5A_44 DNA high_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203186
E10	PvTCF_McCoy's 5A_44	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203167

	DNA high_3					
E10	PvTCF_McCoy's 5A_44 DNA low_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203235
E10	PvTCF_McCoy's 5A_44 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203258
E10	PvTCF_McCoy's 5A_44 DNA low_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203239
E10	PvTCF_McCoy's 5A_72 DNA high_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203221
E10	PvTCF_McCoy's 5A_72 DNA high_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203202
E10	PvTCF_McCoy's 5A_72 DNA high_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203182
E10	PvTCF_McCoy's 5A_72 DNA low_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203217
E10	PvTCF_McCoy's 5A_72 DNA low_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203197
E10	PvTCF_McCoy's 5A_72 DNA low_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203178
E10	PvTCF_McCoy's 5A_72 DNA mid_1	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203246
E10	PvTCF_McCoy's 5A_72 DNA mid_2	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203261
E10	PvTCF_McCoy's 5A_72 DNA mid_3	Asexual_Mixed	PvTCF	Three bio replicates	Brazil _ Acre	SRR10203242

Table 8.1 – RNAseq read files of *P. vivax* and which were combined for *de novo* transcriptome assembly.

Files with the same name in the 'Concatenated Files' column were all combined before assembly construction, with 'Rationale' detailing the reasons why these samples were concatenated together. If the 'Concatenated Files' cell is blank for a given read sample, then that read file was used alone for assembly construction.