UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

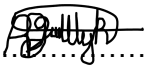
Machine learning and network-based approaches unveil key aspects of gene regulation that drive *Plasmodium falciparum* pathogenesis

by

Roelof van Wyk

Submitted in the partial fulfilment of the requirements for the degree
Philosophiae Doctor in Biochemistry
In the Faculty of Natural and Agricultural Sciences
Department of Biochemistry
University of Pretoria
Pretoria
South Africa

I, **Roelof Daniël Jacobus van Wyk**, declare that the thesis, which I hereby submit for the degree ***Philosophiae Doctor*** in the Department of Biochemistry, at the University of Pretoria, is my own work and has not previously been submitted by me for the degree at this or any other tertiary institution.

SIGNATURE……  ……

DATE……2022/02/07……..

PLAGIARISM DECLARATION

University of Pretoria

Faculty of Natural and Agricultural Science

Department of Biochemistry

Full names of student: Roelof Daniël Jacobus van Wyk
Student number: 28037325
Title of work: Machine learning and network-based approaches unveil key aspects of gene regulation that drive *Plasmodium falciparum* pathogenesis

Declaration

1. I understand what plagiarism is and I am aware of the University's policy in this regard.
2. I declare that this thesis is my own original work. Where other people's work has been used (either from a printed source, internet or any other source), due acknowledgement was given and reference was made according to departmental requirements.
3. I did not make use of another student's previous work and submit it as my own.
4. I did not allow and will not allow anyone to copy my work with the intention of presenting it as his or her own work.

SIGNATURE STUDENT:……..

DATE: 2022/02/22……...

Acknowledgments

I would like to acknowledge the National Research Foundation for awarding me with and NRF Innovation PhD bursary during 2016-2018. I would also like to thank my supervisor Prof. Lyn-Marié Birkholtz for her advice and counsel during the course of my PhD and affording me the opportunity to attend research events abroad. Furthermore, I would like to thank Dr. Riëtte van Biljon for her continued assistance, input and guidance during my PhD. I would also like to thank Prof. Fourie Joubert and Mr. Johann Swart at the Centre for Bioinformatics and Computational Biology for their support and use of the infrastructure provided for analysis of RNA-seq data, GRN construction and hosting of a web-based application of our design. I would also want to acknowledge Dr. Julian Rayner and Dr. Lia Chappell for their support and training with RNA-seq data as well as the assistance on sequencing our sample material.

Lastly, I would like to acknowledge my friends and family for their continued support during the ups and downs of PhD, my mother Fransien van Wyk, my friends Riëtte, Sidhika Hariparsad and Hilde von Grüning.

Summary

Malaria parasites cause human disease through completing a complicated life cycle within both human and mosquito hosts. These organisms are also characterized by numerous molecular eccentricities that make them of immediate biological interest to study. However, the complexity of the parasite life cycle and the composition of its genes and proteins makes studying gene regulation in *Plasmodium falciparum* parasites a multifaceted problem and challenging to resolve.

This doctoral thesis presents the following approaches to study gene regulation using an array of different tools to construct Gene Regulatory Networks (GRNs) for various phases of *P. falciparum* parasite development. 1) We investigated gene regulation of the intraerythrocytic phases of the parasite life cycle, the asexual proliferative phase in which causes the symptoms of malaria as well as the sexual differentiative phase that forms transmissive gametocytes. Initially we investigated the two developmental phases in isolation using time course-based experiments and analysing the data with Dynamic Bayesian Network (DBN) tools. We studied asexual gene regulation using a strategic cell cycle arrest and re-entry experiment, whereby regulatory candidate genes were inferred based on re-entry expression patterns. Application of DBN time course analysis yielded a calcium signalling cascade along with multiple regulatory elements. This approach was expanded to study the sexual development phase as well, using a transcriptomics dataset capturing the daily maturation of gametocytes, which focused on the role of transcription factors. The application of DBN analysis to gametocyte microarray data produced insights into the potential regulatory roles of key ApiAP2 transcription factors which presented with a cascade-like expression as well as putative repressor ApiAP2's which potentially drive the active repression of proliferation-associated transcription.

2) The two developmental phases were also evaluated collectively using RNA-seq datasets sourced from prior research as well as a newly generated gametocyte maturation dataset, capturing all stages of gametocyte development. Integration of data and constructing of a co-expression network lead to a gametocyte associated subnetwork which highlighted potentially novel and significant regulators of gametocyte maturation. The co-expression network itself also constitutes a solid set of curated, cross-dataset normalized genes that can be further used to predict stage-specificity of transcripts in asexual stages of development. Investigations into long non-coding RNA (lncRNA) and their role during gametocyte development was also a key focus of the study. Novel lncRNA were uncovered for gametocyte stages and co-expression network analysis has highlighted many targets of the lncRNA. Investigation into

the role of anti-sense RNA (asRNA) has yielded 9 clusters illustrating the potential for numerous genes (n=285) to be silenced/controlled by their own asRNA.

3) The analysis was further expanded through the construction of a large-scale supervised gene regulatory network using advance ensemble machine learning techniques (GRNBoost2), which evaluated 124 regulatory genes against a total of 5163 target genes. This approached showed great improvements over the previous strategies. This supervised approach was packaged in a user-friendly web application called MALBoost. This application allows user to submit their own transcriptomic data and regulator gene list to perform a choice of two analyses, GRNBoost2 or GENIE3. This approach removes the coding element from the analysis and makes this level of GRN based work available to non-computational biologists.

This thesis presents an in-depth analysis using high-level machine learning and statistical analysis applied to teasing apart the biological significance of transcriptional data. This contributes to our understanding of transcriptional regulation in sexual differentiation and promotes the use of machine learning algorithms in better understanding *P. falciparum* transcriptomes.

Table of Contents

Chapter 1 14

Literature review 14

1.1 Malaria: a global disease burden 14

1.2 Malaria parasites and their complex life cycle 15

1.3 Transcriptional regulation in *P. falciparum* 19

1.4 Gene Regulatory Networks: A general introduction and use case in transcriptomics 24

1.4.1 Correlation networks (CN) 26

1.4.2 Mutual information-based networks (MI) 27

1.4.3 Bayesian networks and Dynamic Bayesian networks (BNs and DBNs) 28

1.4.4 Gaussian graphical models (GGMs) 30

1.4.5 Ensemble methods 31
Random Forest trees (RFs) 32
Gradient Boosting Machines (GBMs) 34

1.4.6 Graph neural networks (GNNs) 35

1.5 Examples of GRN research and their advantages 36

1.6 GRNs in malaria research 37

1.7 Hypothesis 39

1.9 Aim 39

1.10 Objectives 39

Chapter 2 41

Gene regulatory networks describe potential regulatory candidates controlling *P. falciparum* proliferative and differentiative processes. 41

2.1 Introduction 41

2.2 Methods: 43

2.2.1 Gene regulatory network of *P. falciparum* asexual proliferation control points 43
2.2.1.1 Parasite culturing, experimental cell cycle arrest and sampling 43
2.2.1.2 GRN construction from asexual proliferation control points 44

2.2.2 Gene regulatory network of *P. falciparum* sexual differentiation 45
2.2.2.1 Parasite culturing and sampling 45
2.2.2.2 Gene association network for gametocytogenesis 46
2.2.2.3 GRN construction for sexual differentiation 46

2.3 Results: 47

2.3.1 Candidates for cell cycle regulation in *P. falciparum* asexual proliferation inferred as a result of strategic arrest and re-entry 47

2.3.2 The molecular landscape undergoes specific changes as gametocytes progress through development 53

2.3.3 Transcriptional dynamics of sex-specificity and translationally repressed genes 58

2.3.4 Hierarchical contribution of transcription factors and kinases to gametocyte development 59

2.3.5 Resolution of ApiAP2 transcription factors driving gametocyte maturation 61

2.4 Discussion: 63

Chapter 3 68

New insights into the sexual transcriptome of *Plasmodium falciparum* through high-resolution amplification-free RNA-seq and comprehensive gene regulatory network analysis 68

3.1 Introduction 68

3.2 Methods 70

3.2.1 Parasite culture, sampling, and RNA isolation 70

3.2.2 Directional, amplification-free RNA-sequencing 71
3.2.2.1 DAFT-seq library 71
3.2.2.2 DAFT-seq primary data analysis 71

3.2.3 Custom exploration of the gametocyte transcriptome 71
3.2.3.1 Whole transcriptome clustering and comparisons with established gametocyte datasets: 71
3.2.3.2 Weighted co-expression network analysis and intramodular hub gene identification 72
3.2.3.3 Module assignment to stage category data 73

3.2.4 Stage assignment of genes with generalized linear modelling: 73

3.2.5 Text mining and filtering of regulatory machinery from current annotations 74

3.2.6 Evaluation of intergenic long non-coding RNA (lncRNA) and adjacent neighbouring gene pairs through co-expression data 74

3.3 Results 75

3.3.1 High resolution transcriptome from DAFT-seq exhibit clear phase differentiation in the parasite development: 75
3.3.1.1 RNA-seq read quality and mapping 75
3.3.1.2 Sample stage composition and sequence distribution 76
3.3.1.3 Gametocyte signatures present strongly in the sampled transcriptome 78

3.3.2 Normalisation approaches across numerous datasets yield variance stabilising transformation (VST) as the most appropriate 81

3.3.3 Correlation of the total dataset samples yield clear stage categories for general use and assignment of clusters 82

3.3.4 WGCNA highlights a strong bimodal distribution of co-expressed genes for sexual and asexual development in P. falciparum 84

3.3.5 Generalized linear modelling (GLM) reveals predictors of stage predominant genes 92

3.3.6 Subnetwork construction of gametocyte related modules, highlight potentially key gene regulatory elements required in gametocyte maturation 96

3.3.8 Putative specific transcriptional regulators co-cluster in separate early/late expression clusters 100

3.3.9 Adjacent neighbour gene pairs show positive correlations which could indicate shared promoters **101**

3.3.10 Intergenic lncRNA show both strong co- and anti-correlation with genes: **104**

3.3.11 Intragenic lncRNA (anti-sense transcripts asRNA) potentially repress transcripts during gametocyte stages **108**

3.4 Discussion **111**

Chapter 4 **113**

The application of Gene Regulatory Networks in *P. falciparum* through inference-based machine learning approaches. **113**

4.1 Introduction **113**

4.2 Methods: **115**

4.2.1 Consolidate gene candidates for construction of a global GRN in *P. falciparum* **115**

4.2.2 Global GRN for *P. falciparum* using GRNBoost2 **115**

4.2.3 Motif discovery of candidate genes **115**

4.2.4 Web-based application for user friendly access to Gene Regulatory Networks, MALBoost **116**

4.3 Results **117**

4.3.1 Global GRN constructed through supervised machine learning escapes the trappings of conventional correlations and size constraints of Bayesian networks **120**

4.3.2 The distribution of interaction strength differs greatly between candidates **122**

4.3.3 A Focused investigation of select candidates illustrates the ability of GRNs to postulate causal relationships **124**

4.3.4 MALBoost: a web-based application for Gene Regulatory Network Analysis in *Plasmodium falciparum* **128**

4.3.4.1 Intended use 128

4.3.4.2 MALBoost architecture and interface 129

4.3.4.3 MALBoost usage in GRN construction for transcriptional regulator 132

4.4 Discussion **134**

Chapter 5 **139**

Concluding discussion **139**

List of Figures

Figure 1. 1: *Plasmodium falciparum* life cycle in human host. 17
Figure 1. 2: Transcription and regulation in *P. falciparum*. 21
Figure 1. 3: Proposed mechanisms of lncRNA gene silencing in *var* genes for *P. falciparum*. 23
Figure 1. 4: *Gdv1*, HP1 silencing interplay of AP2-G. 24
Figure 1. 5: Gene Regulatory Networks, a general overview. 25
Figure 1. 6: Correlation networks, basic overview. 26
Figure 1. 7: Basic overview of DBNs. 29
Figure 1. 8: The principial differences between GGM and CN. 31
Figure 1. 9: Random Forest tree algorithm. 33
Figure 1. 10: Supervised learning approach as applied in RFs GRN construction. 34
Figure 1. 11: DeepWalk learning process: general overview. 36

Figure 2. 1: *P. falciparum* cell cycle regulation at the G1/S transition point. 48
Figure 2. 2: Molecular mechanisms controlling cell cycle re-entry. 49
Figure 2. 3: A GRN generated by GRENITS on differentially expressed genes driving cell cycle re-entry in asexual *P. falciparum* parasites. 51
Figure 2. 4: Gene interaction network analysis reveals the molecular landscape of gametocyte maturation. 57
Figure 2. 5: Specific regulatory elements contribute towards gametocytogenesis. 61
Figure 2. 6: Expression of ApiAP2 during gametocyte development. 62
Figure 2. 7: Specific regulatory elements contribute towards gametocytogenesis. 63

Figure 3. 1: RNA-seq quality statistics. 76
Figure 3. 2: Inclusion of high-resolution gametocyte development transcriptome through RNA-seq based platforms. 78
Figure 3. 3: Gametocyte associated genes reveals expected expression trends throughout the time course data. 79
Figure 3. 4: Meerstein-Kessel gold standard of sexual development genes for the in-house RNA-seq dataset. 80
Figure 3. 5: Dataset normalisation strategies. 82
Figure 3. 6: Sample severances through Pearson correlations classified within developmental blocks. 83
Figure 3. 7: Weighted co-expression network analysis captures a bimodal distribution of co-expression in developmental stages. 87
Figure 3. 8: Gene ontology (GO) for network modules capture through pie scatter plots and module density shown with ellipses. 90
Figure 3. 9: Gold standard asexual and sexual gene sets from Meerstein-Kessel compared to co-expression network: 91
Figure 3. 10: Generalised linear model output of stage-associated genes. 93
Figure 3. 11: Meerstein-Kessel gold standard cross-referenced with GLM staged genes and intramodular hub genes. 94
Figure 3. 12: ROC output for stage-associated gene panel from *GradientBoostingClassifier*. 95
Figure 3. 13: Subnetwork constructed from gametocyte associated modules and text-mining. 96
Figure 3. 14: Subnetwork from gametocyte-associated modules. 97
Figure 3. 15: Neighbouring gene pairs captured using *csuWGCNA*. 102
Figure 3. 16: Intergenic lncRNA captured using *csuWGCNA*. 105
Figure 3. 17: Subnetwork of lncRNAs show correlated and anti-correlated expression with gametocyte modules. 107
Figure 3. 18: Intragenic anti-sense RNA (asRNA) for in-house time course dataset (day -2 to 13). 110

Figure 4. 1: Consolidation of candidate genes for transcriptional regulation in *P. falciparum*. 119
Figure 4. 2: Distribution of inferred regulatory interactions from GRNBoost2. 121
Figure 4. 3: Ranked distribution of top interactions per candidate gene, prioritising candidates. 123
Figure 4. 4: Gene Regulatory Network analysis captures strong association with candidate genes and their respective targets. 126
Figure 4. 5: Expression profiling across the intra-erythrocytic development cycle reflect findings from GRN. 127
Figure 4. 6: MALBoost user overview. 129
Figure 4. 7: Internal architecture of MALBoost web-based application. 130
Figure 4. 8: MALBoost web-based application for GRN construction in *P. falciparum*. 132
Figure 4. 9: MALBoost results for AP2-G GRN. 134

List of Tables

Table 3. 1: Datasets used for co-expression network with a total of 49 samples through all datasets. 81
Table 3. 2: Associated statistical breakdown showing the number of connections per gene involved in the subnetwork in Figure 3.15 98

List of abbreviations

AP2/ERF:	Apetala2/Environmental Response Factors
asRNA:	antisense RNA
BNs:	Bayesian networks
CDS:	coding-domain sequence
CNs:	correlation networks
CSP:	circumsporozoite protein
DAFT-seq:	directional amplification-free RNA-seq
DAGs:	directed and acyclic graphs
DBNs:	dynamic Bayesian networks
DMFO:	difluoromethylornithine
G:	gametocyte
GAN:	Gene Association Network
GBMs:	Gradient Boosting Machines
GDV1:	gametocyte development protein 1
GGM:	Gaussian Graphical Models
GLM:	generalized linear modelling
GNNs:	Graph Neural Networks
GO:	gene ontology
GRNs:	Gene Regulatory Networks
GVS:	Gibbs Variable Selection
HAT:	histone acetyl transferase
HDAC:	histone deacetylase
IDC:	intraerythrocytic developmental cycle
LLINs:	long-lasting insecticide-treated bed nets
lncRNA:	long ncRNA
lncRNA-TARE:	telomere-associated long non-coding RNAs
malERA:	Malaria Eradication Agenda
MC:	Monte Carlo simulation
MI:	mutual information networks
Mix:	mixed gametocyte and asexual
MMV:	Medicines for Malaria Venture
ncRNA:	noncoding RNA
PCC:	Pearson correlation coefficients
PTM:	post-translational modifications
RAM:	Random Access Memory
RE:	re-enter
RFs:	Random Forest trees
RT:	ring-trophozoite
SCC:	Spearman correlation coefficients
scRNA-seq:	single cell RNA-seq
TF:	transcription factors
TFBS:	transcription factor binding sites
TOM:	topological overlap similarity matrix
TS:	trophozoite-schizont
TSS:	transcription start site
UTR:	untranslated region
VST:	variance stabilizing transformation
WGCNA:	weighted gene co-expression network analysis

Chapter 1

Literature review

1.1 Malaria: a global disease burden

Malaria devastates the global health and economy annually, causing an estimated total of 229 million cases and 409 000 mortalities during 2019 alone[1]. With global case numbers rising from 217 million cases in 2014 to 229 mil, there is cause for alarm. African countries are the most burdened by this disease accounting for an estimated 215 million cases in 2019. Mortality rates are also highest amongst African countries, with ~94% of mortalities from this region in 2019[1]. Most disturbing of all is that children under the age of 5 account for ~67% of the deaths. These numbers also account for millions of cases that have been averted through efforts by the World Health Organization with programs such as Malaria Eradication Agenda (malERA) and the Roll Back Malaria initiative[1–3]. Together these agencies aim to reduce malaria mortality and case load by 90% in 2030 by tackling malaria transmission on several fronts[4].

The primary method for controlling malaria transmission is targeting the obligate intermediate mosquito vector. The malaria parasite is carried to humans by the females of multiple *Anopheles* mosquito species, making vector control, chiefly through the use of insecticides, an integral part of a successful elimination strategy. Previous efforts almost solely focused on indoor residual spraying and formed the primary focus of the Global Malaria Eradication Program (GMEP, 1955–1969). However, the current tool kit for targeting vector transmission focuses on co-implementation of indoor residual spraying and the use of long-lasting insecticide-treated bed nets (LLINs)[4]. This strategy forms the basis of combatting widespread resistance of the *Anopheles* vectors to the four major classes of commercially available insecticides: pyrethroids, organochlorides, organophosphates and carbamates[5]. In addition, partnerships with endemic countries aim to increase ecological surveillance of mosquito breeding areas and general insecticide resistance[4].

Furthermore, malaria elimination also requires targeting the parasites that cause the disease. Currently, there is a critical need for new approved mechanisms (chemotherapies and vaccines) that combat critical attributes of the parasite including the development of rapid resistance to monotherapies and insidious evasion of the human immune response. The first malaria vaccine approved for use by the WHO, the RTS, S vaccine uses a truncated form of an essential parasite antigen, the circumsporozoite protein (CSP) coupled to adjuvants to

produce an immune response protecting against future infections. This vaccine offers partial protection against malaria in African children, preventing an estimated 40% of malaria cases in three African countries[6]. Further optimization of this promising vaccine candidate has shown promise, with the second iteration of this vaccine, R21, showing around 77% efficacy in a small-scale study in Burkina Faso[2]. However, as the vaccine is not completely effective and has multiple challenges for mass-distribution i.e. storage, accessibility, compliance with 4-dose regimen, treatment with anti-malaria drugs remains a mainstay of intervention against malaria infection.

The most widely used antimalarial chemotherapies, quinolone and artemisinin derivates, arose from ethnobotanic knowledge dating back hundreds of years[3]. These medicines target essential functions of the parasite. Chloroquine and quinolone derivatives prevent the parasite from sequestering toxic heme moieties that form through the parasite's digestion of human haemoglobin, while artemisinin is activated through the release of ferric iron during this same process, enabling the drug to alkylate numerous parasite proteins, causing acute toxicity to the parasite[4]. A third compound class targeting folate biosynthesis in the parasite is comprised of sulphadoxine and pyrimethamine and are typically paired with an artemisinin-derived partner drug for treatment. However, for all of these compound classes, encroaching drug resistance has been reported first in Southeast Asia and recently in Africa, necessitating rapid development of novel antimalarial drugs[7].

To combat the virulent spread of malaria drug resistance, researchers aim to provide a continual pipeline of antimalarial drugs in development, which has resulted in several frontrunner compounds with novel targets entering clinical development i.e. KAF156, DSM256, MMV253 and MMV048[5] among others. These compounds represent efforts to adhere to guidelines formulated by the Medicines for Malaria Venture (MMV, www.mmv.org) to produce long lasting, effective treatments against malaria that can produce single dose cures, prevent transmission of the parasite and/or combat resistance. These leaps in progress against fighting malaria were also enabled by uncovering key elements in understanding the malaria parasite's complicated biology, a process which is still largely incomplete.

1.2 Malaria parasites and their complex life cycle

Malaria in humans is cause by five species of *Plasmodium* parasite, *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi* of which *P. falciparum* accounts for the most severe form of the disease[1]. It's also the most widespread form with the highest number of infections

and mortalities attributed to *P. falciparum* globally. *P. falciparum* is the focus of this study and will be discussed in more detail.

P. falciparum infections in humans are initiated by the introduction of sporozoites as the result of a female *Anopheles* mosquito blood meal, where the parasites are injected through the skin and make their way to the liver through the blood stream of the human host[6]. The sporozoites will transiently pass-through hepatocytes in the liver through the formation of non-replicative vacuoles before finally productively invading the cells. A parasitophorous vacuole houses the sporozoite in the hepatocyte while it divides asexually into a hepatic schizont and bursts open to release invasive merozoites for initiation of the erythrocytic invasive cycle (Figure 1.1).

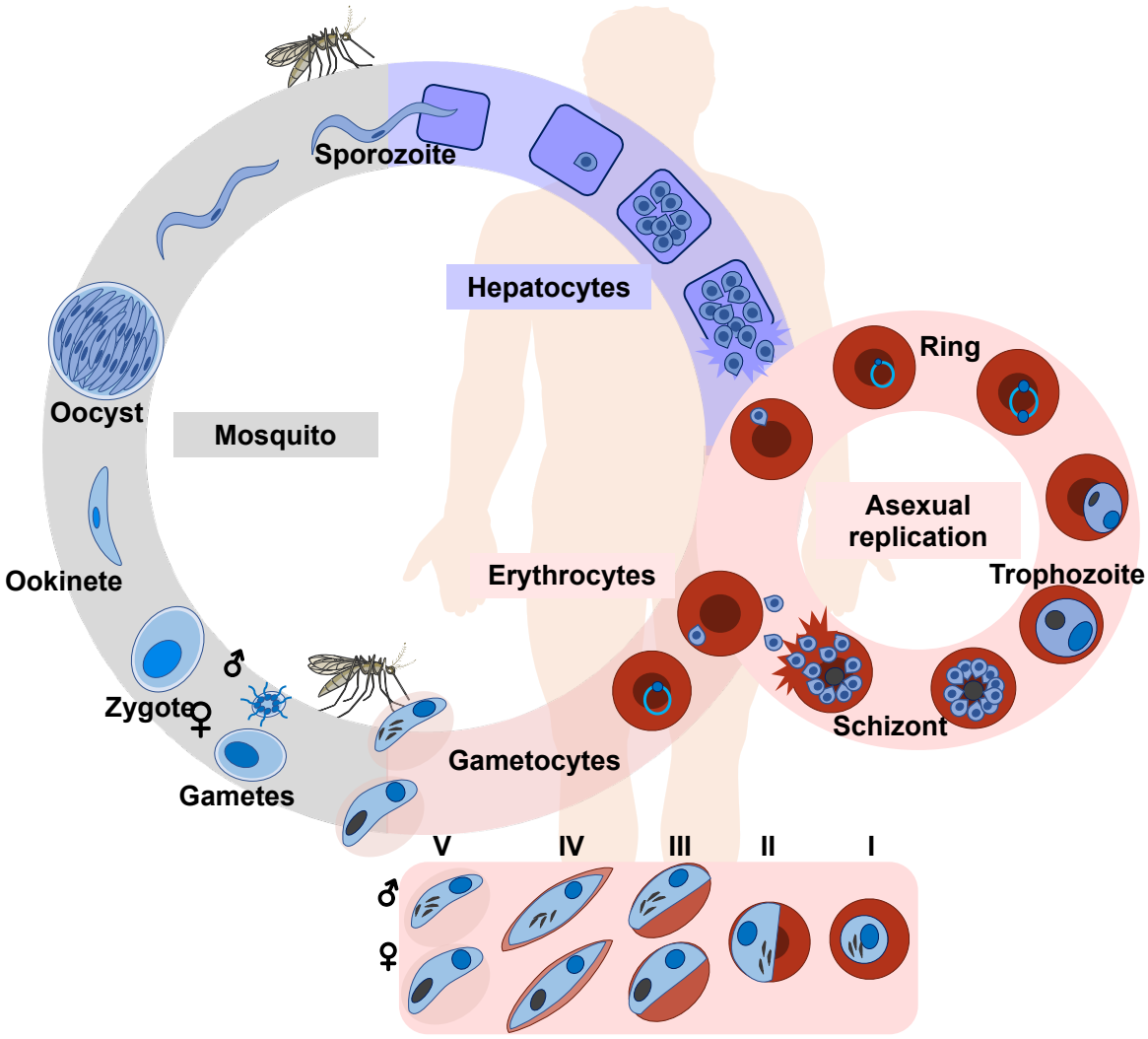


Figure 1.1: *Plasmodium falciparum* life cycle in human host.
The parasite life cycle is a complex system ranging from sexual development in the mosquito vector to three human host phases. During the human host phases the parasite starts its development in liver hepatocytes, then moving on to erythrocytes where parasites proliferate over a ~48hr period into schizonts, which after bursting releases ring parasites to repeat the process. A small subset of ring parasites commits to sexual development which occurs in the bone marrow and spleen and has five distinct stages. Gametocytes develop either as male or female. These gametocytes when mature re-enter circulation and find their way to the mosquito vector via a blood meal. Produced using Servier medical art under creative commons license, provided with permission by Dr. Riëtte van Biljon.

Merozoites invade erythrocytes through apical organelles, encasing themselves in a parasitophorous vacuolar membrane and initiating the intraerythrocytic developmental cycle (IDC). Once invaded, the parasite enters the relatively inert morphological ring-stage for its first 15 hours. The parasite then undergoes a substantial increase in metabolic activity as it enters the trophozoite stage. Trophozoites digest the haemoglobin of their host cells, sequestering toxic heme moieties into hemozoin crystals visible in the parasite and increasing in size and exporting proteins that permeabilise the erythrocyte membrane to metabolites essential to the parasite's survival[8]. In the final hours of the 48 hour IDC, the parasite enters

the schizont stage, during which the parasite replicates its DNA and divides into 20-30 daughter merozoites[7], ready to reinitiate the cycle.

A small proportion (<10%) of asexual parasites stochastically commit to sexual development in each cycle, resulting in a prolonged gametocyte maturation process of 10-12 days in *P. falciparum*. The parasite progresses through five morphological stages of gametocyte development to produce sexually dimorphic, mature gametocytes that can transmit to the *Anopheles* mosquito vector[9,10]. This commitment event occurs either within the ring-stage, that then develop into stage I gametocytes (same-cycle conversion) or in schizonts, that develop into stage I gametocytes after the merozoites reinvade (next cycle conversion)[11]. Stage I gametocytes are morphologically indistinguishable from trophozoites but express numerous sexual-stage specific surface proteins[12] and *in vivo* are sequestered in the bone marrow of the human host[13]. From stage II of gametocyte development, the parasite begin to elongate through the construction of a subpellicular membrane network[14], aiding in keeping the parasite sequestered in the bone marrow. The early stages of gametocytes also continue to digest haemoglobin in the red blood cell, but hemozoin crystals are more diffuse than those visible in asexual trophozoites (Figure 1.1). The male and female gametocytes also become distinguishable as the female have more dense hemozoin crystals compared to the males with hemozoin crystals scattered through their cytoplasm[15]. Stage IV gametocytes are the most elongated and pointed, with females clearly distinguishable by the presence of osmiophilic bodies and the parasite ceases haemoglobin digestion at this stage[16]. As the parasite matures into stage V gametocytes, the subpellicular microtubular network depolymerizes, resulting in sausage-shaped gametocytes with rounded ends. *In vivo,* stage V gametocytes are the only stage of sexual development that circulates in the bloodstream and after a few days become infectious to feeding *Anopheles* mosquitoes[17].

Once gametocytes are taken up into the mosquito, they rapidly activate to produce microgametes through exflagellation of male gametocytes and macrogametes from female gametocytes[18]. Microgametes and macrogametes fuse into a zygote, which develops into a motile ookinete that migrates through the mosquito midgut before productively invading a midgut cell. The ookinete forms an immotile oocyst containing numerous sporozoites that burst out and travel to the mosquito's salivary glands, ready to reinitiate the infectious cycle[19].

1.3 Transcriptional regulation in *P. falciparum*

The complicated life cycle development of *P. falciparum* is underpinned by a highly regulated cascade of transcription. The transcriptional profile of the IDC is particularly illustrative of the fine regulation the parasite can implement[20]. During this development phase, nearly all the parasite transcripts are expressed in a stage-specific manner. Gametocyte development characterized by the expression of a number of gametocyte specific transcripts[21] and sex specific transcripts[22]. General transcriptional control elements for *P. falciparum* remain conserved with that of most eukaryotes, with most members of the transcription initiation complex identified[23]. However, *cis*-acting promoter-proximal elements, i.e. TATA-boxes remain challenging to identify in the AT-rich (~80%) *P. falciparum* genome despite the availability of numerous transcription start site (TSS) datasets[24–26].

Many genes have multiple TSS sites, which are preferentially used at different points in the TSS. These TSS differences were associated with a marked difference in gene expression level, which points to chromatin availability as a major regulator of promoter strength[24]. Discrepancies for mean differences were noted ranging from as little as 7 bp to 496 bp which was attributed to thresholding difference between studies[25]. Approximately 81% of TSS blocks occur within 1000 bp of the start codon and a further 65% of these occur less than 500 bp of CDS[24]. Leaderless transcription (lacking a 5'UTR) with TSS within exons for both single-exon and multi-exon genes were also observed. This produces a complex view of transcription in *P. falciparum*, were the TSS may vary greatly for certain genes and often lack 5'UTRs complicating the data further. Studies on TSS have also found a convincing bias for a TA dinucleotide (as is usually found in eukaryotes at the +1 position) at the TSS during asexual development, with the presence of two GC-rich regions approximately 150 bp and 210 bp downstream of the TSS[24].

Chromatin structure and accessibility has been shown to be a major determinant of overall gene expression in *P. falciparum*[27,28]. Evaluation of chromatin structure and accessibility for regulatory events, were uncovered for of genes. The temporal accessibility of genes to their nearest local regulatory elements correlates for the majority (~85%) of genes[29]. This does not exclude the existence of distant regulatory enhancer regions, however it seems likely that the majority of genes are regulated by their nearest set of regulatory elements[29]. In fact distant regulatory sequences identified were mainly relegated to studies in centromeres, ribosomal DNA loci and subtelomeric regions[29,30]. Epigenetic factors also act as chromatin modellers and interpret the chromatin environment to regulate the concerted expression of genes. Open and closed chromatin environments regulate access of transcription factors (TF) to gene

promoter regions which allow these *trans*-factors to act as transcriptional repressors or activators[31]. The combined evaluation of nucleosomal landscape with activating or repressive histone-tail modifications were able to predict overall transcript levels in the *P. falciparum* transcriptome to an accuracy of ~80%[29], showing that the chromatin epigenetic landscape is a critical aspect of gene regulation in the parasite. The resultant landscape is exceptionally open in the IDC, largely contributed by the absence of the H1 linker histone[32] and an abundance of activating epigenetic marks[33]. During this state, transcription factor binding sites (TFBS) are exposed and transcription can be initiated (Figure 1.2). These same types of in-depth analyses have not been performed on the sexual stages of the blood phase of malaria infection, and it is largely unknown how the chromatin landscape changes in gametocytes to allow for its differentiated transcriptome.

The *P. falciparum* genome contains a relatively small (~1% of total genes) cadre of putatively predicted sequence-specific transcription factors mostly consisting of helix-turn-helix and zinc-finger domain containing proteins[34]. The largest and most extensively investigated family of TFs in *P. falciparum* parasites comprises 27 proteins containing Apicomplexan AP2 (ApiAP2) DNA binding domains, homologous to the Apetala2/Environmental Response Factors (AP2/ERF) family of proteins found in *Arabidopsis thaliana*. The ApiAP2 proteins that have been investigated for their effect on transcription so far have been found to be either repressive[35] or activating in nature[36–39] (Figure 1.2). Of the 27 ApiAP2's, 21 have been paired with a specific DNA motif *in vitro* and in the rodent malaria species, most ApiAP2 proteins were associated with the passage of the parasite through key points in its lifecycle, most of these occurring in the mosquito[40]. However, in the human malaria parasite *P. falciparum*, one factor has been shown to be essential for completion of the IDC, AP2-I[41], AP2-G has also been highlighted as the key gatekeeper to sexual commitment[42,43], while loss of the repressive transcription factor AP2-G2 prevents the completion of gametocyte maturation and AP2-HS specifically regulates the parasite's heat shock response, suggesting that at least some of these proteins have specific roles in regulating transcription in the parasite[35,44,45]. In addition, some of these proteins SIP2 and AP2-Tel have been shown to specifically associate with heterochromatin or telomeres and do not have a marked influence on gene expression[31,46]. While evidence from DNA pull-downs with specific DNA sequence implicate the ApiAP2 family of transcriptions factors as the only sequence-specific binding factors in the parasite[29], at least one other transcription factor (MYB1) has been shown to bind specific DNA sequences *in vitro* and affect the ability of the *P. falciparum* to progress through the IDC[47].

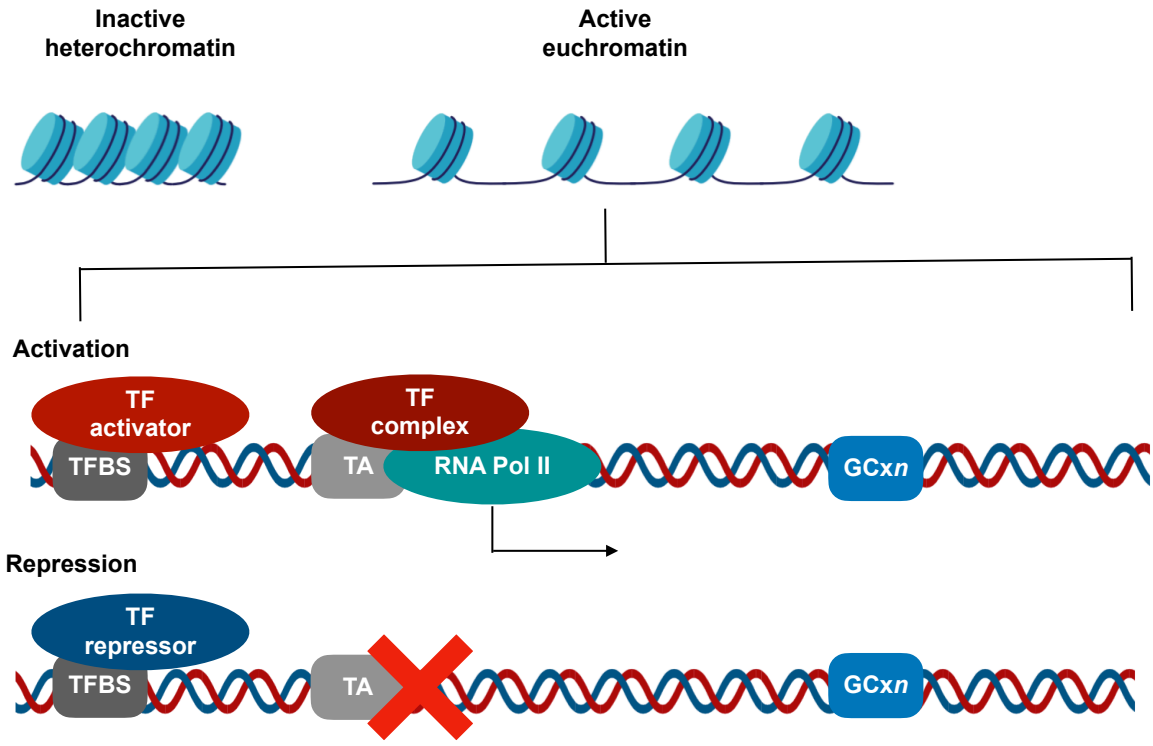


Figure 1.2: Transcription and regulation in *P. falciparum*.
Inactive heterochromatin state reduces access to promoter regions, thus preventing transcription. Active euchromatin provides access of promoters for TFs which may either lead to activation or repression of transcripts. During activation the TF recruit TF complex proteins and RNA Pol II which initiates transcription. TA: dinucleotide TSS. GCx*n*: GC-rich regions. TFBS: transcription factor binding site, TF: transcription factor.

While transcriptional regulation is a major factor in regulation of gene expression in *P. falciparum* parasites, it has also been shown that the parasite employs post-transcriptional controls such as alternative splicing, post-transcriptional modifications, mRNA stabilization[48,49] and mRNA decay[50]. However, RNA interference is presumed to be missing as no known homologue for DICER exists in the parasite genome[51]. Capping of pre-mRNA, polyadenylation and splicing occur in the parasite nucleus where alternative splicing may produce transcripts for translational repression[22]. Alternative splicing also plays a role in regulating variant gene switching, which is important for immune evasion and isoform production. The transcripts exported from the nucleus can also be subject to further regulation instead of being translated. Some transcripts are stabilized and translationally repressed by large protein complexes such as DOZI-CITH complex and the pumillo-family (Puf) proteins[52,53], with specific relevance in female gametocyte development in preparation for gametogenesis. Other transcripts which are not protected by stabilizing proteins are quickly degraded. This process occurs through the activity of ribonucleases and decapping enzymes (DCP1/2) degrade mRNA in a 5'-3' direction. Alternatively, 3'-5' degradation is mediated by the CCR4-NOT complex and CCR4-

NOT associated factor 1 (CAF1), which mediates deadenylation[54] and transcript degradation is performed by the exosome.

Post-translational control can also affect transcriptional regulation, predominantly as a result of the modification of histones. These post-translational modifications (PTMs) are primarily the addition and removal of acetyl and methyl groups[32]. Acetyl modifications are deposited by histone acetyl transferases (HATs) and its removal is facilitated by histone deacetylases (HDACs), with bromodomain-containing proteins which often act as readers of these modifications. Methylation modifications to histones can result in mono-, di- and tri-methylation of lysine residues and mono-, or di-methylated arginine residues[55]. Histone lysine methyltransferases such as SETs deposit methyl groups on lysine residues, while histone arginine methyltransferases such as HRMTs modify arginine residues[56]. Demethylases for each of these groups are responsible for the removal of methyl groups. The contribution of histone PTMs to gene regulation is clearly established and associated with stage-specific usage of particular PTMs in a dynamic fashion[57–60]. This contributes to the overall euchromatic nature of asexual parasites and a repressive, more heterochromatic environment that characterises gametocyte differentiation[57–60].

In addition to the above, noncoding RNA (ncRNA) are transcribed RNAs that do not yield protein products but are considered to be an essential component of the transcriptome[42]. A specific group of ncRNAs which are interesting in *P. falciparum,* are long ncRNA (lncRNA). These lncRNA play specific roles in gene silencing of the strongly regulated *var* gene family in IDC parasites, which only allow for the expression of a single exported protein in a specific parasite[43,61]. These subtelomeric *var* genes contain a variable exon 1 and a conserved exon 2 and two lncRNA is said to interfere with their expression. These lncRNA are transcribed from a bidirectional promoter housed between the two exons which create a lncRNA that is antisense (complimentary) to the gene and a sense strand which runs along the 2nd exon (Figure 1.3A). Both these lncRNAs are capped but not polyadenylated and are transcribed by RNA Pol II[42]. This complex formation leads to gene silencing of *var* genes, particularly those located near the telomeres[40,61]. The full mechanism remains to be resolved, however, there is a proposal that ncRNAs can lead to site-specific histone modifications[42]. Two such modifications are known to silence *var* genes: H3K36me3 and H3K9me[62,63]. A SET protein (PfSETvs), which performs the H3K36me3 modification, has been shown to be recruited to *var* genes through Pol II[62]. This creates the possibility that lncRNAs could be indirectly recruiting histone modifiers in gene silencing for these virulence genes[42,63]. This is bolstered by the role of lncRNA-mediated nucleosome positioning in other organisms[42]. Members of the *var* gene family which are internally located on chromosomes, appear to be activated by their

intron-derived lncRNA counterparts, illustrating the complexity of lncRNA in regulation of genes[64].

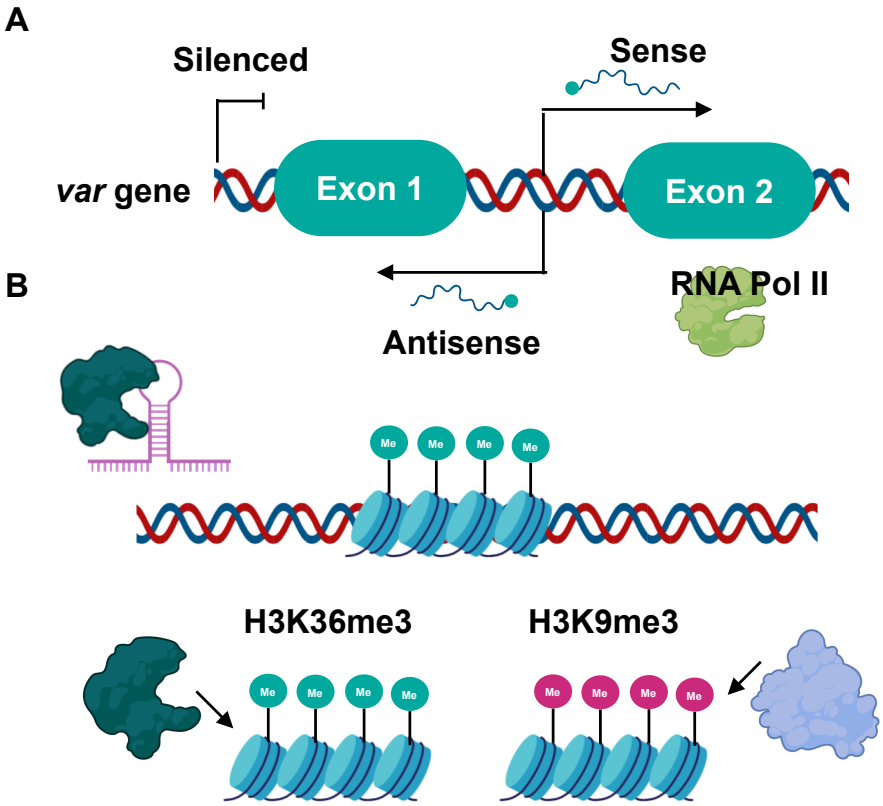


Figure 1.3: Proposed mechanisms of lncRNA gene silencing in *var* genes for *P. falciparum*. A) Two lncRNA are transcribed at a bidirectional promoter situated between exon 1 and 2 by RNA Pol II. One antisense strand (complementary to the transcript) and one sense strand (running along the gene) is transcribed. B) lncRNA and RNA Pol II complex leads to the potential recruitment of histone modifiers such as PfSETvs. Histone modifications known to silence *var* genes are H3K36me3 and H3K9me3. Adapted from[42].

In addition to lncRNA residing in the introns of genes, regulation of specific genes by antisense RNA (asRNA) have also been observed. One such mechanism affects gametocyte development protein 1 (GDV1), asRNA silences *gdv1* and prohibits sexual development of the parasite in human host stages[45]. GDV1 plays an essential role in evicting a repressor protein heterochromatin protein 1 (HP1), which continuously silences AP2-G, a transcription factor responsible for initiating gametocyte development during ring-stage parasites. With the presence of GDV1, HP1 is no longer bound upstream of AP2-G and the AP2-G product can initiate sexual differentiation (Figure 1.4). The negative regulation for GDV1 asRNA could occur at the level of mRNA transcription, stability or translation[42,45]. It's clear from these data that ncRNA can play an essential role in gene regulation, though the research on these mechanisms in *P. falciparum* is still extremely sparse.

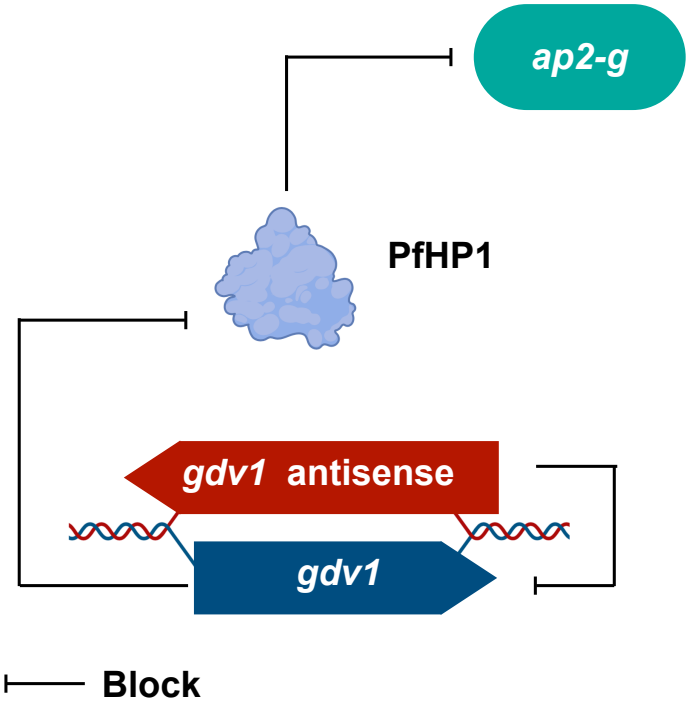


Figure 1.4: *Gdv1*, HP1 silencing interplay of AP2-g.
Gdv1 asRNA serves to silence *gdv1* which in turn cannot evict HP1 from the upstream location of AP2-G. Adapted from[42].

1.4 Gene Regulatory Networks: A general introduction and use case in transcriptomics

Gene Regulatory Networks (GRNs) describe the relationship between regulatory genes and their respective target genes[65,66]. GRNs are useful to characterise the regulatory genes or control agents required during developmental processes. In its simplest form, a GRN connects a "regulatory" gene (also referred to as candidate gene) to its target gene, capturing the effector *vs*. affected relationship known as edges (Figure 1.5). Several methods for modelling or simulating these relationships exist that all use high-throughput experimental data. The use of expression profiles, genomic sequences and transcription factor binding arrays provide essential data for such modelling[67]. Several methods are available and some of the most common approaches and methods of particular interest to this study include correlation networks (CNs) and mutual information networks (MI), Bayesian networks (BNs) and dynamic Bayesian networks (DBNs), Gaussian Graphical Models (GGM), Ensemble learning methods such as Random Forest trees (RFs) and Gradient Boosting Machines (GBMs), and Graph Neural Networks (GNNs) (Figure 1.5). Each of these will be briefly introduced below.

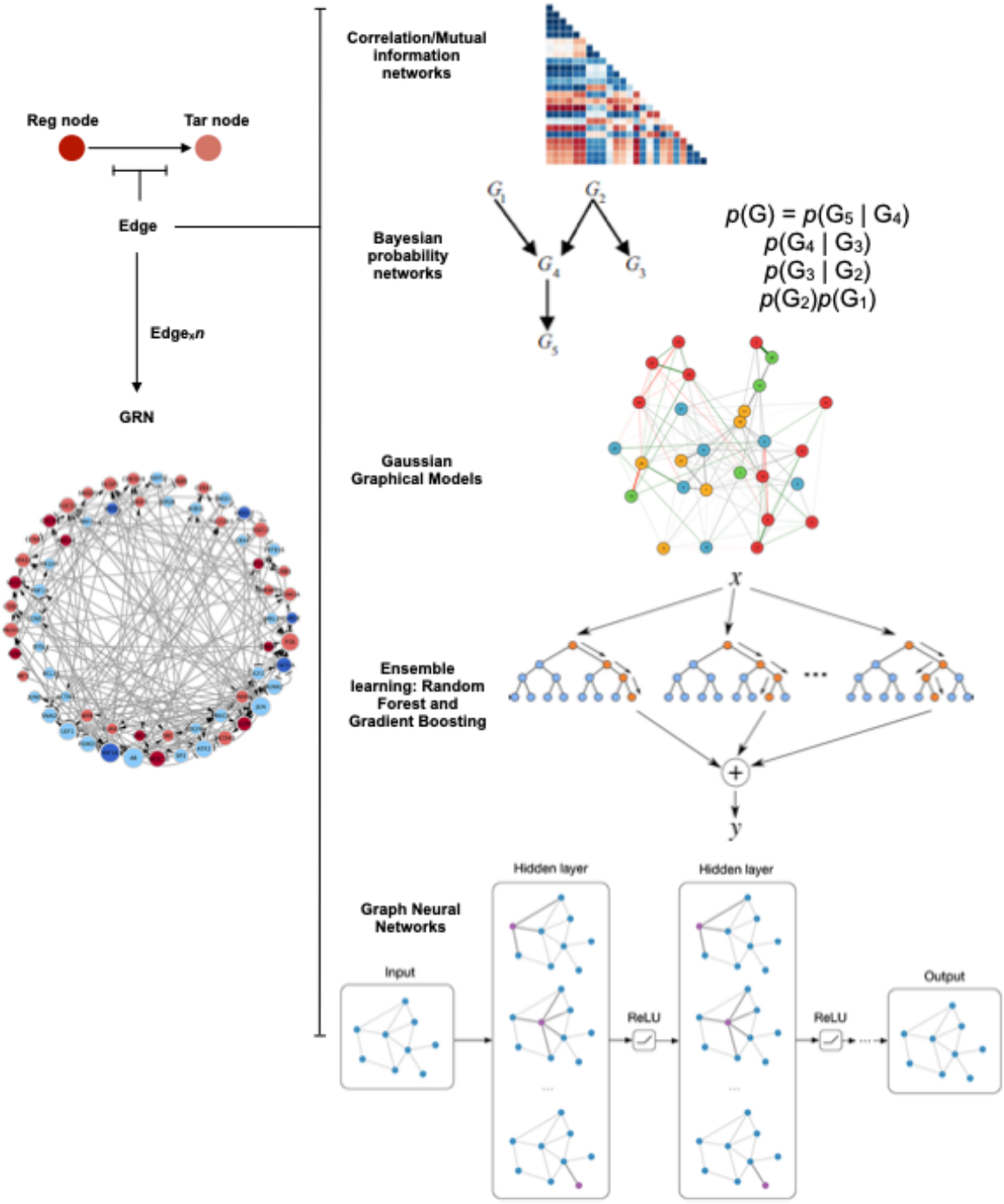


Figure 1.5: Gene Regulatory Networks, a general overview.
GRNs describe the relationships between two nodes (Reg node: regulatory gene or Tar node: target gene) in terms of edges (the value that explains the relationship). These edges may be derived in a number of ways, the most commonly used methods are, correlation networks (also known as co-expression networks) and mutual information networks (usually a knowledge-based relationship), Bayesian networks such as dynamic Bayesian networks, Gaussian Graphical Models, Ensemble learning methods such as Random Forest trees and Graph Neural Networks. This collection of relationship explanations constitutes the overall GRN.

1.4.1 Correlation networks (CN)

CNs are achieved by calculating correlation coefficients via a pairwise assessment of genes in relation to all other genes[68]. From these relationships, a full gene interaction network is constructed with the correlation values between genes used to describe the edges of the network. Generally, a form of module definition is then applied such as hierarchical clustering (Figure 1.6), such as with weighted gene co-expression network analysis (WGCNA)[69]. As an example of these module formations, the stronger correlated nodes shown in red (Figure 1.6) would form the theoretical module due to their stronger relationship and association.

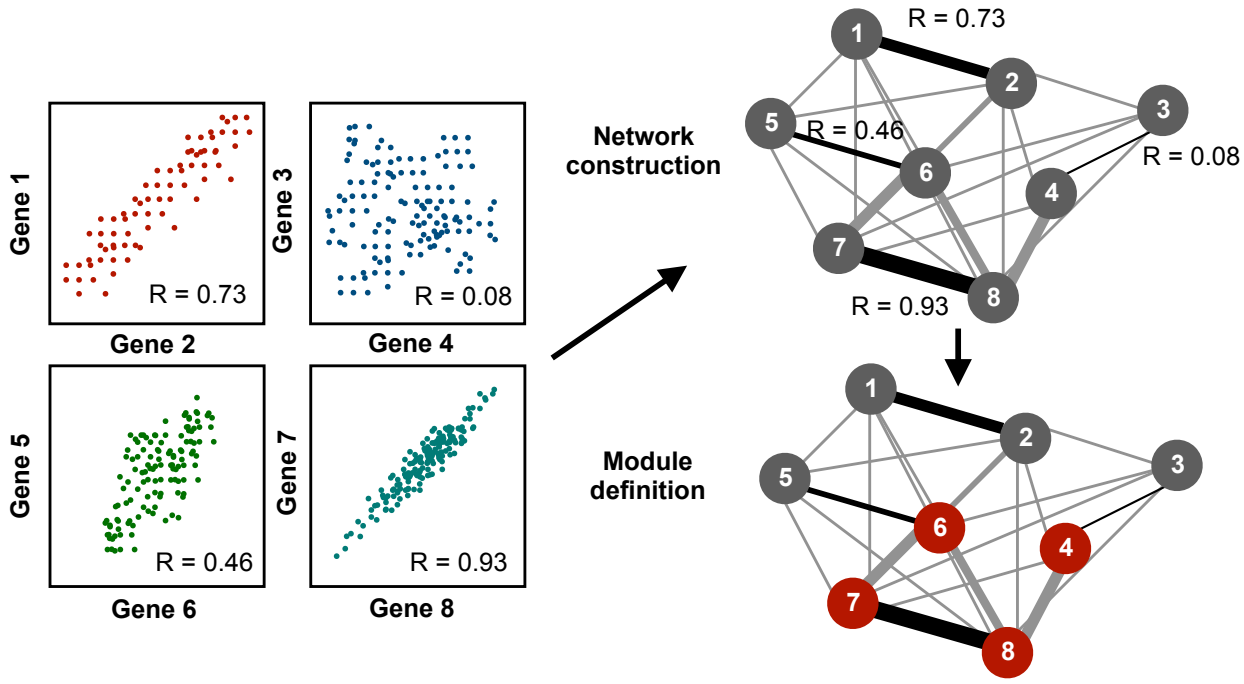


Figure 1.6: Correlation networks, basic overview.
Correlations are calculated between different gene pairs, the quantified relationship (edge) is used to construct the full network. After this, module definition is applied to extract different modules within the network based on their relationships. Red nodes in the network shows the application of module definition where these strongly correlated nodes form a particular module or cluster.

Various different methods for calculating coefficients exist: Pearson correlation coefficients (PCC), which best describe linear relationships within a dataset and Spearman correlation coefficients (SCC) which allow for non-linear relationships[68]. CNs as a rule of thumb require as little as 16 samples to perform reliable calculations, however, an increase in sample size will invariably lead to an increase in statistical power. Many software packages (R-based) have been developed to perform these calculations and may apply an array of statistical tests to measure the significance of the relationships between genes. Packages such as CoXpress are popular choices in the field[68] and use PCC and hierarchical clustering in order to determine

differentially correlated modules between samples or conditions. The use of clustering implies that arbitrary cut-offs are required to form modules, which could lead to reproducibility issues in the analysis. DiffCorr, which is another popular R package, uses PCC and Fisher Z-transformations to identify differentially correlated sets and is mostly used for experimental condition comparisons. Other tools such as Molecular Complex Detection incorporate the use of GO analysis, but have become rather dated[70]. These networks remain largely unweighted correlation networks, which without proper statistical evaluations, could wrongly emphasize several unrelated and randomly correlated relationships[69].

This caveat between the difference in simple Pearson/Spearman correlations as compared to weighted correlation networks such as WGCNA, lies in escaping the trapping of biologically coincidental occurrences through establishment of scale-free network topology[69]. Scale-free topology is a network whose degree distribution follows power laws, this is to say that characteristics of the network are independent of the size of the network. Networks of this nature retain a relatively constant underlying structure when the number of nodes increase. WGCNA achieves this scale-free topology through the application of a power coefficient as determined through a function that evaluates the scale-free nature of the network at various power coefficients. An appropriate threshold is applied here to ensure that the network holds to this topology dynamic. Chiefly WGCNA proves beneficial through the use of topological overlaps and hierarchical clustering in modules identification, as such the use of hard thresholds is often not necessary[70]. These analyses produce useful clusters/modules of correlated gene sets, while being able to assign a strength of relationship measure which holds true for scale-free topology assumptions, proving to be a powerful tool in GRN construction. Correlation coefficient networks often struggle at defining more complex statistical dependencies such as non-linear relationships[71].

1.4.2 Mutual information-based networks (MI)

MI measures the amount of information that a given random variable contains about another. Uncertainty about one variable can therefore be reduced based on the knowledge of another variable[72]. In practical terms informative gene feature components are predictive of the expression outcome of the target gene as exampled by a network database configuration such as STRINGdb. MI share a feature with CNs in that they are sensitive to indirect regulatory relationships which can lead to stochastic edges in the network. The remedy for this feature is the use of conditional mutual information, which rely on the mutual information of a third variable in any given pair comparison.

1.4.3 Bayesian networks and Dynamic Bayesian networks (BNs and DBNs)

BNs remain popular choices for GRN construction with Bayes probability theorem at its core. These probabilities are combined with graph theory for modelling of the GRN. These graphs are directed and acyclic graphs (DAGs) with a set of local probability distributions[73]. This relationship is described as $G = (X, A)$, where X represents the nodes/genes and A the direct edges/probabilities of the graph. BNs decompose node sets into conditionally independent node subsets which are non-overlapping with regards to their position. Here, the algorithm works through a process of discovering the best DAG configurations through belief propagation. A key problem with BNs is the lack of cyclic inferences in the DAGs, this means that configurations such as A→B→A is not possible in the inference. These graphs have direction and do not have cycles or feedback loops[74], the lack of which is often solved through the process of Dynamic Bayesian models using time series data[71].

DBNs are an extension of BNs that infers interaction uncertainties through the use of probabilistic graph models[73]. This allows DBNs to infer cyclic interactions which are important in biological networks as feedback loops may exist. DBNs try to resolve optimal DBN structures from given gene expression data. Once the structure has been reconstructed, a set of probabilities can be learned from methods such as maximum likelihood. Structure learning for DBNs can involve any number of methods such as local search, stochastic global optimisation or Monte Carlo simulation (MC)[75]. Compared to BNs which are static, DBNs uses conditional probabilities based on discrete time increments[76]. The simultaneous modelling of timescales account for faster interactions within time increments and slower interactions which occur much earlier or later in the events record. The models almost construct a sequential order which makes variables in the next time increment dependent on the previous increment (Figure 1.7). DBNs are commonly used in GRN construction, although the technique has become dated with the advancement of ensemble learning and deep learning methods.

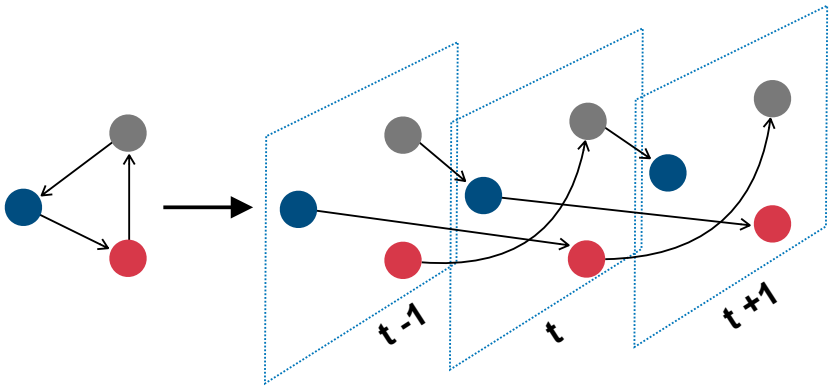


Figure 1.7: Basic overview of DBNs.
DBNs use timeseries data and account for the conditional probabilities between time increments. This infers the relationship between faster interactions (occurring within a single time step) with that of slower interactions (occurring either in the beginning or later in the series).

1.4.4 Gaussian graphical models (GGMs)

GGMs are a group of probabilistic models which assume that gene expression are jointly Gaussian distributed and represent conditional dependencies between genes as undirected graphs[71]. Here, a correlation network is constructed and transformed into a partial correlation network (an undirected graph) and these are known as GGMs[77]. In order to infer the directed nature of GGMs, the graph is then converted into a partially directed graph through estimating the pairwise ordering of nodes from multiple testing of the log-ratios of standardized partial variances. These partial orderings when projected onto the GGM infer the underlying causal network as a subgraph of the directed edges[77] (Figure 1.8). The algorithm removes edges from the independence graph to obtain the underlying DAG. Limitations of these models lie with the assumption of Gaussianity in the distribution, which goes hand in hand with linear dependencies between variables. GGMs are undirected and require heuristics to infer direction from the graphs, which may prove to be challenging[71].

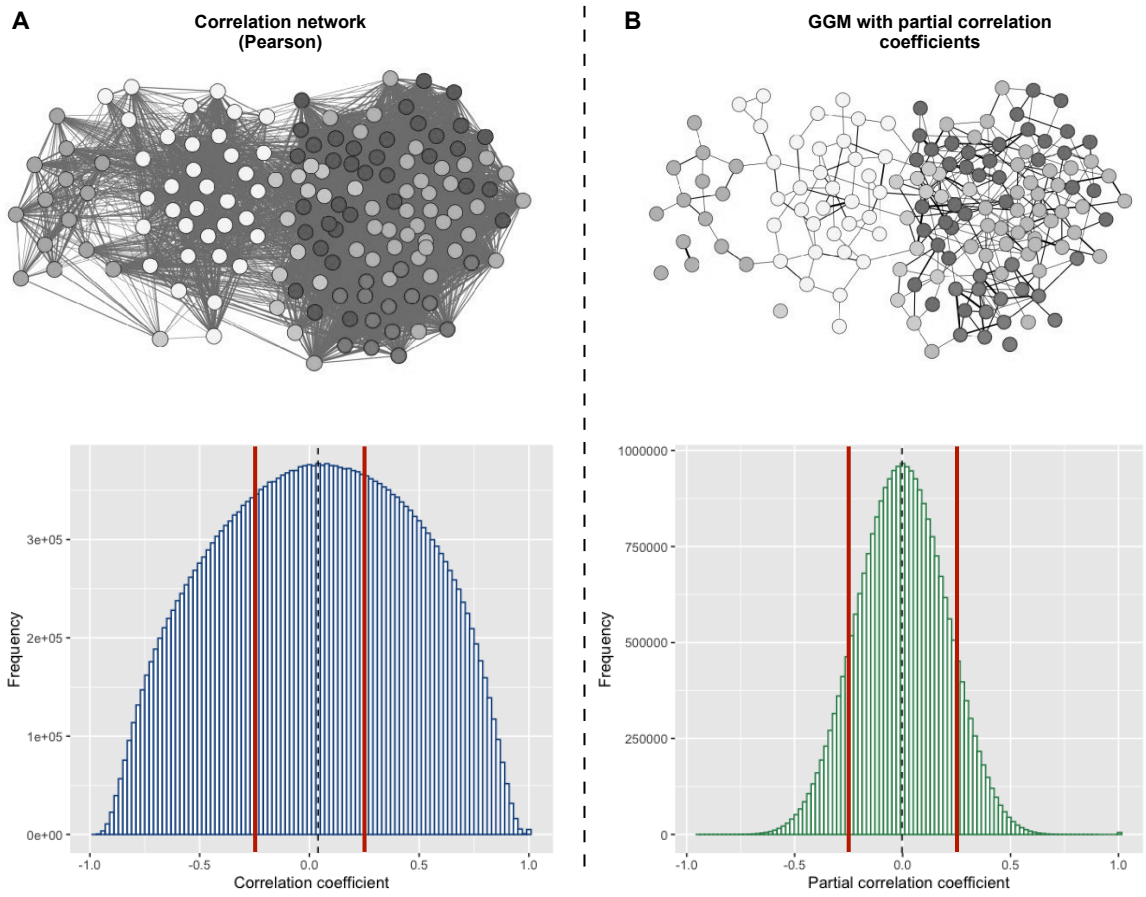


Figure 1.8: The principial differences between GGM and CN.
A) CNs construct all possible correlation coefficients for all genes with a large distribution. B) GGM uses partial correlation coefficients (correlates which match a Gaussian distribution) to form a sparse network with a more centred median. Median indicated with dashed line; red lines indicate significance regions. Distribution plots were constructed from microarray data of IDC development to illustrate real world examples[49].

1.4.5 Ensemble methods

Ensemble methods loosely define processes that make use of multiple models to solve a particular research question, hence the name ensemble. Specifically, ensemble models constitute a family of machine learning algorithms that typically construct several models during the training process and reach a consensus for all the models constructed. The most popular algorithms associated with this class of models are RFs and GBM. Each of these two classes also contain various algorithm forms within them and so the general modular basis will be discussed with a few specific examples of application.

Random Forest trees (RFs)

One of the most prominent examples of the use of RFs in GRN construction is with an algorithm called GENIE3, which was the best performer of the DREAM4 challenge in 2010[71]. RFs effectively employ decision trees at its core, which splits the data based on feature values causing bifurcation at each step. The result resembles that of a tree with the data split into several branches. The drawback to a singular decision tree is that it often works well only on the observed data but struggles to fit new data and thus can produce inaccurate predictions. RFs solve this problem while utilising the simplicity of decision tree methods. RFs perform random sampling of the given data through a process of bootstrapping. Bootstrapping is just a method of randomly sampling from data with the one key feature being that data points may be reselected. This implies an iterative process whereby the data is randomly sampled with each step. A subset of features in each step is also selected rather than all the features and a decision tree is constructed for this step. This means that the rout nodes of each decision tree are constructed from randomly sampled data and randomly selected features to produce a tree. This process is repeated often thousands of times and thus multiple trees describing the data is conceived. Each tree is assessed for its output in terms of prediction and the consensus of all the trees are taken to answer the prediction, this is referred to as bagging. From the randomly sampled data points, some which were not selected (out-of-bag) are then used to assess the trees for accuracy, almost like a test set would be used. Each step of forest construction can vary the number of features used in tree construction and ultimately the best forest is used which describe the data best based on out-of-bag assessment (Figure 1.9).

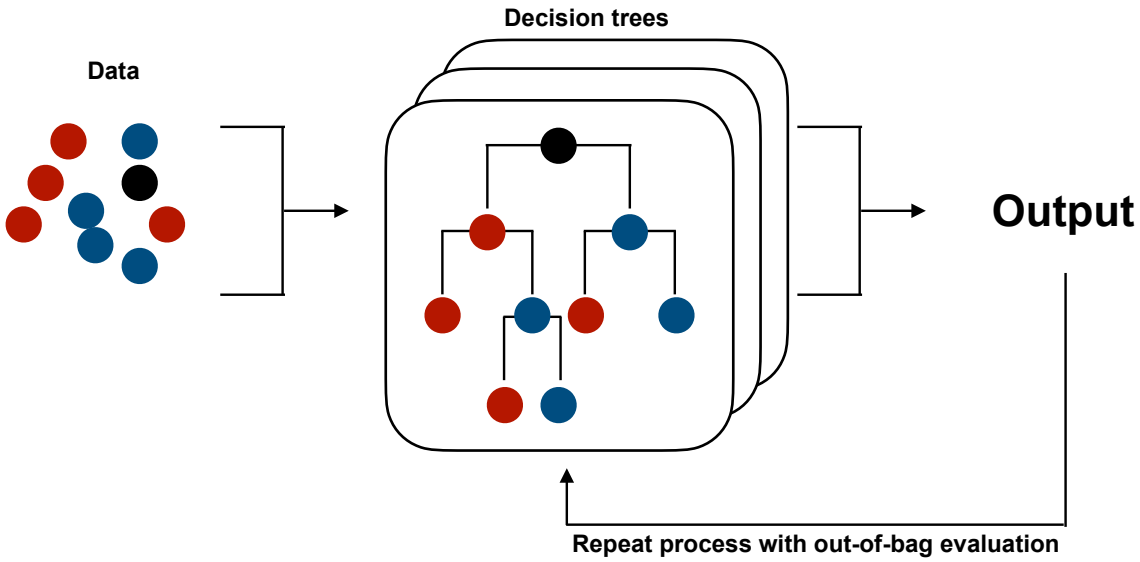


Figure 1.9: Random Forest tree algorithm.
Input data and their features are randomly sampled through bootstrapping and numerous decision trees are constructed based on these samplings. Each tree is evaluated for a consensus output on the out-of-bag (unsampled data) which describes the overall accuracy of the forest. The process repeats several times and the forest with the highest accuracy is chosen for the model.

The application of RFs in constructing GRNs use a set of features (usually the expression values of multiple transcription factors or known regulatory genes) in a repetition to predict the values of target genes. Simply put, the data are split into potential regulatory genes from which the model feature set is to be constructed and potential target genes, which are all the genes of the data in an iterative way. The ranked importance of each regulatory gene/feature is extracted from the final forest model which quantifies each features contribution to the predicted target gene values. The feature which was most instrumental in explaining the predicted outcome therefore has a greater probability of regulated the target gene (figure 1.10). These ranked importance values describe a directed network[71,78,79]. This essentially describes the basis of the GENIE3 algorithm which infers regulatory gene importance with regards to each gene in the dataset.

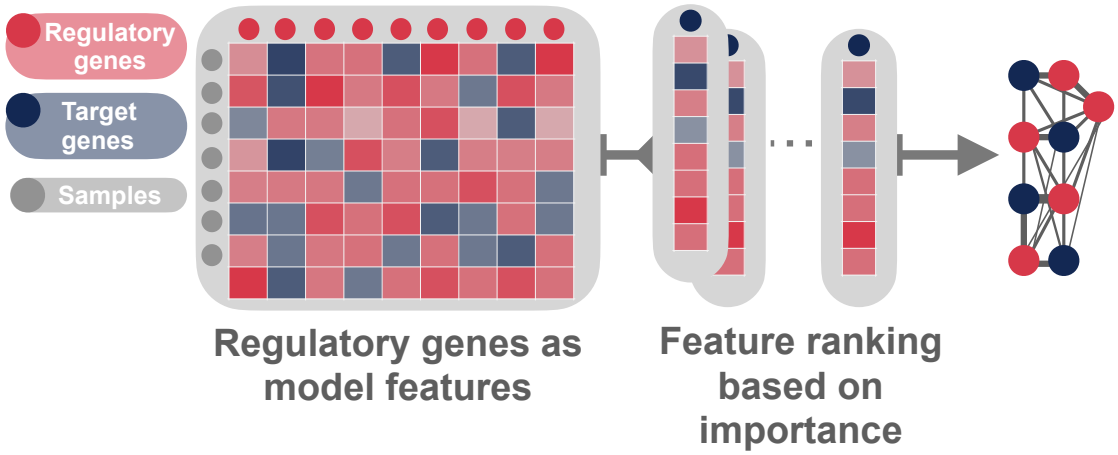


Figure 1.10: Supervised learning approach as applied in RFs GRN construction.
Selected regulatory genes formulate the key variables in the matrix from which the RFs are constructed and each gene in the dataset is predicted. The features are ranked for the relative importance in each target gene prediction and a directed network is constructed from this.

Gradient Boosting Machines (GBMs)

GBMs are similar to RFs with the major difference inherent in how the algorithm learns form decision trees. GBMs are used in GRN construction chiefly in the same manner as RFs. A popular tools for such network construction is GRNBoost2, which forms part of the Arboreto suite of GRN tools[78]. This tool was evaluated in the same DREAM5 challenge as GENIE3, and outperformed GENIE3 while reducing computation time.

GBMs learn from short/small trees of a set size and uses the errors from the previous tree to inform the next tree, rather than constructing complete trees and consolidating them as is done with RFs. Initially the algorithm starts of by using an assumed answer to explain the output values, such as an average of the values. Then the residual for each observation is calculated and small trees of user-defined size is constructed from the features to explain these residuals. The output leaves of these trees are used to recalculate residuals for the data and another round of tree construction is performed. Using a learning rate (a set constant), which helps to reduce overfit in the data, the residuals and new averages are scaled for each tree. The successive trees will then get closer to the observed data through multiple rounds of this process with evaluation at every step to help improve the model learning. This algorithm thus attempts to get the residuals as close to zero as possible. This describes a process that aims to essentially perform multiple small steps in the right direction, which ultimately increases accuracy in low variance datasets commonly seen in test set data. This algorithm drastically reduces the time with which the model can calculate the accurate outcomes and

reduces variance[78]. The residuals can be calculated a number of ways and the function which perform this calculation is referred to as the loss function[80].

1.4.6 Graph neural networks (GNNs)

GNNs are a class of deep learning algorithms commonly used in the GRN construction. This class of algorithm contains specialised methods which convert graph information into vectors which the neural network can learn from. Numerous sub-architectures for this class exist, with graph embedding techniques (GET)[81] being a sub-type relevant to GRN research. Neural networks much like GRNs can be described in terms of nodes and edges which explain the particular graph. Typically, the process starts with and adjacency matrix, these matrices explain the edge values between nodes and captures the topology for graphs. When the graph is unweighted and undirected then edges values will be set at 1. Additional values for nodes are stored in separate matrices as they will contain attribute information regarding nodes such as gene expression. The use of this adjacency matrix suggests a prior knowledge of gene interactions, this can be based on physical evidence such as protein-protein interactions, TF-binding arrays, metabolic networks or even correlations.

The goal of these techniques is to predict what other gene pairs are likely to interact. DeepWalk[82] uses Word2Vec frameworks which are well known for their use in natural language processing which learning embeddings by performing multiple random walks for each node of the graph and then optimises through Skipgram objective function[81]. Skipgram learns embeddings for a node such that it maximises the probability of predicting each related node in the random walk. This process is similar to that of another tool GNE which uses random sampling of interactions rather than random walk approaches such as Word2vec[83]. These steps just serve to illustrate some form of vectorisation of the given graph data, many of forms of vectorisation exist each with their own complexity in constructing vectors from the graph. The algorithm uses these vectors in the Skipgram function to learn from nodes in their local proximity and produce output embeddings (Figure 1.11). The separation of these data post-embeddings is illustrated by the red and blue node separation.

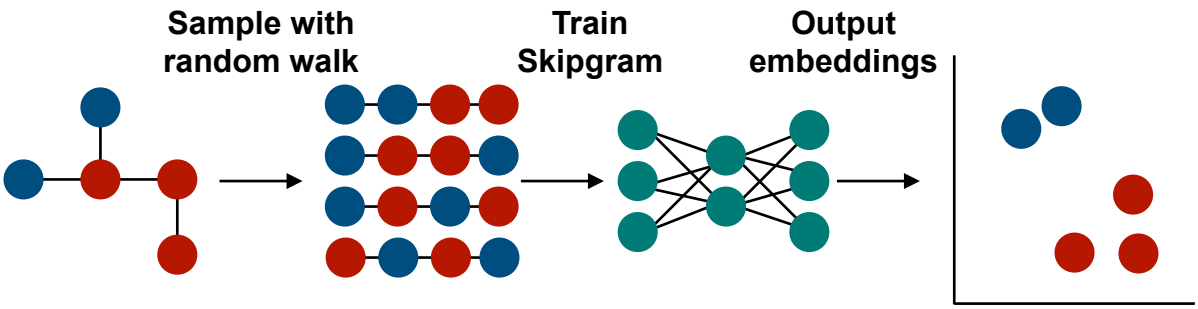


Figure 1.11: DeepWalk learning process: general overview.
Graph networks are randomly sampled through random walk algorithms and then vectorised for training in Skipgram. Skipgram performs local learning of the nodes and outputs embeddings (in green shows the neural network) which describe the underlying relationships for interaction networks. The data are spatially separated according to the relationships from the embeddings show a clear separation of red and blude nodes effectively forming clusters

1.5 Examples of GRN research and their advantages

The use of GRNs is ubiquitous across molecular research in various cell types and disease interests, perhaps none more so than cancer research. The underlying structure of these networks can produce powerful insights into relevant gene discovery, reprogramming mechanisms, drug target identification, clinical outcome predictions and many more[84]. As illustration of the usefulness of GRNs in biological research, the discovery of relevant genes involved in oligodendrogliomas (type of brain tumour) with a 1p/19q chromosomal co-deletion, was achieved through the construction of a GRN from gene expression data as well as copy number data[85]. Here, eight genes with strong implications on signalling pathways and 14 genes with implications on metabolic pathways were resolved from this network. A well-known *ELTD1* glioblastoma oncogene showed overexpression while a tumour suppressor gene (*SLC17A7*) under-expressed. This information obtained from the GRN analyses therefore provided clear mechanisms ascribed to the disease pathology in these cell types. However, the power of GRNs could also be extended to evaluate epigenetic alterations that may possibly enhance epigenetic reprogramming of paragangliomas[85].

This form of transcriptional reprogramming inference through GRNs analysis is an important task. The application of GGMs have proven advantageous in this area of cancer research, particularly applied in ovarian cancers and different platinum responses, and rewiring between breast cancers of luminal A subtype and basal-like subtype[86]. These differential networks rediscovered known genes associated with platinum resistance in ovarian cancer. This process utilised existing data from The Cancer Genome Atlas (https://www.cancer.gov/) and some of the existing static GRN data generated. However, more than just resistance can be

investigated for cancers and GRNs have also been used to examine and identify potential new drug targets. For example, a drug target prioritization method that evaluated an array of algorithms such as MI, CN, GGM, RFs and support vector machines (SVMs) were able to capture known interactions and indicate novel interactions to predict druggability of gene products[87]. The latter was used to assess target genes derived from the networks and 75% of their targets were revealed as potential drug targets. BN analysis of breast cancer data yielded two targets for combination therapy, *mTOR* and *STAT3* genes[88]. This led to the experimental implication of cryptotanshinone as a potent modulator resulting of apoptosis in breast cancer cell lines.

Beyond the scope of drug interactions and the identification of potentially novel drug targets, GRNs are also useful in predicting clinical outcomes for certain types of cancer[84]. The use of SyNet and 11 other network frameworks in conjunction with survival-labelled gene expression data was evaluated for breast cancer[89]. SyNet is a synergistic pairwise gene comparison network, built from the survival-labelled data. SyNet is estimated at ~85% accuracy, particularly so for the breast tissue-specific networks.

The scope of GRN analysis and advantages in cancer research has been vast. However, the use of GRNs in malaria research remains relatively limited despite available transcriptomic data for *P. falciparum* and other *Plasmodium* species.

1.6 GRNs in malaria research

The use of inferential GRNs in *P. falciparum* research has been sparse. An integrated network derived from various experimental data showed the potential to resolve genes involved in erythrocyte invasion during the IDC of the parasite[90]. Although the network generated relied on some of the approaches used in GRN construction, it more accurately constitutes a simple interaction or association network and produced little insight into transcriptional regulation. An early study of GRNs in *P. falciparum* research focused on the parasite's IDC using microarray data[91]. Variational Bayesian Expectation Maximization (VBEM) approaches were used as an early attempt to resolve a "skeleton" cell cycle GRN whereby 59 interactions were resolved for 38 putative regulatory candidates[92]. These candidates included proteins such as transcription factor MYB2 and a calcium-dependent protein kinase. This skeleton cell cycle network provided a first draft of how cell cycle control proteins interact to regulate the IDC.

Modelling of the Maurer's cleft pathway through GGM resolved key candidates that aid in the transport of proteins from the parasite cytoplasm to the host erythrocyte surface[66]. The focus, however, was only related to genes in Maurer's cleft and does not extrapolate to general gene regulation in the parasite. More recent work incorporated phenotypic data to understand molecular factors at the heart of low- and high-transmissibility of malaria parasites[93]. Through the use of weighted co-expression networks constructed from large DNA-microarray datasets, strong insights regarding malaria transmission could be provided and highlighted the involvement of AP2-G, histone deacetylase 1 (HDAC1) and a putative histone deacetylase (HDA1)[93].

In this PhD thesis, we resolved and published two critically important GRNs for *P. falciparum*, which will be discussed in greater detail in Chapter 2[94,95]. Both networks were constructed using DBN methods and time series data. With this, we describe two networks focused on different phases of parasite development, 1) the proliferative phase of the parasite during its IDC and 2) during the differentiative phase associated with gametocyte development. The networks allowed inference of regulatory elements controlling proliferation and differentiation. However, these GRNs were created on transcriptome data generated from DNA microarray studies. Whilst these datasets are informative regarding interactions and regulation of particular genes, this data does not lend itself towards more in-depth analyses of additional regulatory mechanisms like lncRNAs. RNA-seq datasets are available for the asexual parasite's IDC and allowed for inferences of intricate regulatory mechanisms[96], this level of resolution was lacking for the prolonged gametocyte development process of *P. falciparum*. We therefore generated a complete RNA-seq dataset spanning the entire gametocytogenesis process for *P. falciparum* (Chapter 3). We could use this information to generate the largest and most comprehensive inferred gene regulatory network in the field, which ultimately lead to the creation of an applied solution to the identification of gene regulatory elements in malaria parasites (Chapter 4).

1.7 Hypothesis

Machine learning and network-based pattern deconvolution can delineate factors involved in transcriptional regulation in various life cycle stages of *P. falciparum*.

1.9 Aim

This study aimed to construct GRNS to provide an integrated overview of transcriptional regulation for both asexual proliferative phases as well as sexual differentiation in *P. falciparum* and provide a tool for researchers to perform these types of analyses without prior knowledge of scripting languages.

1.10 Objectives

1. Construct GRNs to aid in the understanding of factors which drive asexual proliferation and sexual differentiation (Chapter 2)
2. Generate an in-depth transcriptome of all stages of gametocyte development with RNA-seq to investigate the relationship between molecular factors and their mutual information driven process during sexual differentiation (Chapter 3)
3. Utilise the power of machine learning and GRN techniques to evaluate regulatory factors in *P. falciparum*, while offering an applied solution to the field (Chapter 4)

Outputs

Manuscripts:

1. van Biljon, R., Niemand, J., **van Wyk, R.** *et al.* (2018) Inducing controlled cell cycle arrest and re-entry during asexual proliferation of *Plasmodium falciparum* malaria parasites. Sci Rep 8, 16581
2. van Biljon, R., **van Wyk, R.,** Painter, H.J. *et al.* (2019) Hierarchical transcriptional control regulates *Plasmodium falciparum* sexual differentiation. BMC Genomics 20, 920
3. van der Watt ME, Reader J, Churchyard A, Nondaba SH, Lauterbach SB, Niemand J, Abayomi S, van Biljon RA, Connacher JI, **van Wyk RDJ**, Le Manach C, Paquet T, González Cabrera D, Brunschwig C, Theron A, Lozano-Arias S, Rodrigues JFI, Herreros E, Leroy D, Duffy J, Street LJ, Chibale K, Mancama D, Coetzer TL, Birkholtz LM. (2018) Potent *Plasmodium falciparum* gametocytocidal compounds identified by exploring the kinase inhibitor chemical space for dual active antimalarials. J Antimicrob Chemother 1;73(5):1279-1290.

4. van Heerden, A., **van Wyk, R**., Birkholtz, L., (2021) Machine learning uses chemo-transcriptomic profiles to stratify antimalarial compounds with similar mode of action. Cellular and Infection Microbiology. Front. Cell. Infect. Microbiol. 11:558.
5. **van Wyk, R.,** van Biljon, R., Birkholtz, L. (2021) MALBoost: a web-based application for Gene Regulatory Network Analysis in *Plasmodium falciparum.* Malaria Journal. 317.

Conferences:

1. The exploration of the *P. falciparum* gametocyte transcriptome using RNAseq. EMBL Conference BioMalPar XIV: Biology and Pathology of the Malaria Parasite. May 2018. Meyerhofstraße 1, 69117 Heidelberg, Germany. Poster
2. Application of network analysis on *Plasmodium falciparum* transcriptomes at the MRC Office of Malaria Research Conference (MOMR) 31 July- 2 August 2016, Pretoria, South Africa. Poster

Chapter 2

Gene regulatory networks describe potential regulatory candidates controlling *P. falciparum* proliferative and differentiative processes.

2.1 Introduction

The *P. falciparum* life cycle is characterized by numerous developmental bottlenecks, specifically in transitions between the mosquito to human or vice versa that are of interest to study for the potential of transmission blocking interventions. However, there is only one point in the life cycle at which there is a binary option for which developmental path to follow: the point at which blood-stage malaria parasites commit to either asexual (>90%) or sexual (<10%) development. During the asexual replicative cycle of the *P. falciparum* parasite, thousands of merozoites are released in the blood stream to initiate the 48 h IDC, resulting in malaria pathology due to massive expansion of parasite numbers. This phase of development is characterized by a tightly controlled cyclic transcriptional cascade of ~85% of the genome[20,97]. The parasite's cell cycle during the asexual proliferative phase includes peculiar features e.g. asynchronous nuclear divisions within one schizont and specific mechanisms for organelle segregation and morphogenesis of daughter merozoites[98–102]. In unicellular protists like *Plasmodium* parasites, cell cycle control is more closely related to developmental control and there is a divergence from canonical cell cycle regulation features[101]. Conversely, during sexual differentiation, the specific activation of the AP2-G transcription factor[10,103,104] results in a complete reprogramming of the transcriptome as the parasite progresses through terminal differentiation into male or female gametocytes, completely losing the cyclic transcriptional activation pattern of asexual development[95,105].

In terms of specific molecular regulators, several atypical cyclin-dependant kinase (CDKs), CDK-related proteins, cyclins and other CDK regulators[99,100] have been implicated in cell cycle control in *P. falciparum,* although their involvement in the regulation of canonical eukaryotic mitotic cascades is unclear. Despite the identification of some putative regulators of gene expression, including the ApiAP2 family of transcription factors[106–108] and epigenetic regulation of particular gene families[60,62], the specific mechanisms controlling transcriptional activation in the parasite are incompletely understood, with recent data clearly showing mRNA dynamics are also influenced by additional post-transcriptional mechanisms[48,49]. In depth analyses of

unknown regulatory factors associated with cell cycle control during asexual proliferation and gametocytogenesis of *P. falciparum* is needed.

A GRN depicts transcriptional regulation and provides an understanding of the dynamics and interaction of genes and GRNs are powerful tools to characterise the regulatory primers or control factors associated with developmental processes[65,66]. Malaria parasites have a complex life cycle with associated multifaceted biology and regulatory mechanisms. This drive both proliferative processes during asexual replication of the parasite – causing disease pathogenesis; or differentiative processes during gametocytogenesis – ensuring disease transmission. The parasite is able to tightly control is gene expression during both these phases of the life cycle[20,94]. Although certain regulatory elements like AP2 transcription factors have been associated with this regulation, the intricacies of stage-specific gene regulation have not been clarified. This makes the parasites' transcriptome, across both the proliferative and differentiative phases of the life cycle, a perfect case to apply GRN analyses. This should elucidate the dynamics associated with transcriptional regulation and could identify regulatory factors of importance in parasite developmental biology. With more and more data being produced for *P. falciparum*, particularly transcriptomes, a need for deeper interpretation of the data comes into demand.

The use of inferential GRNs in *P. falciparum* research has certainly not been common although the potential for resolving genes involved in erythrocyte invasion during the IDC of the parasite has been indicated with simplistic gene association networks (GANs) like those generated by the STRING database[90] (https://string-db.org/). STRING uses several input data including empirical protein-protein interaction data from e.g. yeast 2-hybrid systems to connect gene pairs to each other. Although BNs often produce impressively accurate results, VBEM could resolve only a preliminary GRN for the *P. falciparum* cell cycle[92]. Such Bayesian methods normally also prove resource intensive and often limits researchers to evaluating a small subset of genes at a time[76,109].

The popularity of BNs provides an intriguing approach for deeper investigations of the parasite's transcriptional regulation. Several advances in computational power and a wider knowledge of transcriptional regulators, creates a climate where BNs may produce greater insights[76]. DBNs in particular have proven to be easily implementable frameworks in our experience for various aspects of *P. falciparum* regulatory networks[94,95].

Here, we utilised new datasets that describe the global transcriptome of malaria parasites at various timepoints during both their proliferative and differentiative phases. Firstly, DNA

microarray-based transcriptome analyses were performed on asexual parasites to evaluate a cell cycle transition point important for proliferation. Secondly, the complete process of conversion of asexual parasites to gametocytes and the whole of gametocyte development was followed with transcriptomics. In both instances, the power of DBNs was harvested to analyse these datasets by using an R-based implementation: Gene Regulation Network Inference Time Series (GRENITS)[109]. GRENITS uses a combination of DBN and Gibbs Variable Selection (GVS) in conjunction with linear modelling to determine sets of interaction probabilities between regulatory nodes and target nodes. In this manner, we were able to construct two distinct GRNs capturing the asexual proliferative and sexual differentiation of the parasite and provided information on regulatory elements thereof[94,95].

The content of this chapter has been published in part in the following instances:

1. van Biljon R, Niemand J, **van Wyk R**, Clark K, Verlinden B, Abrie C, von Gruning H, Smidt W, Smit A, Reader J, Painter H, Llinas M, Doerig C, and L Birkholtz (2018) Inducing controlled cell cycle arrest and re-entry during aseuxal proliferation of *Plasmodium falciparum* parasites. Scientific Reports, 8:16581, doi:10.1038/s41598-018-34964-w (IF 4.609).

2. van Biljon R, **van Wyk R**, Painter HJ, Orchard L, Reader J, Niemand J, Llinás M, Birkholtz L. (2019). Hierarchical transcriptional control regulates *Plasmodium falciparum* sexual differentiation. BMC Genomics, Dec 3;20(1):920.DOI: 10.1186/s12864-019-6322-9, (IF 3.501)

In each instance above, the experimental component was contributed by R van Biljon. The data generated was subsequently used and computational work pertaining to GRNs were performed in this PhD. **The computational data analysis in these papers were therefore driven by the PhD candidate R van Wyk**.

2.2 Methods:

2.2.1 Gene regulatory network of *P. falciparum* asexual proliferation control points

2.2.1.1 Parasite culturing, experimental cell cycle arrest and sampling

In vitro cultivation of intraerythrocytic *P. falciparum* parasites held ethics approval from the University of Pretoria (EC120821-077). A cDNA microarray dataset was generated on the Agilent G2600D Microarray Scanner (Agilent Technologies, USA) for parasite samples that were produced following cell cycle arrest induced in asexual parasites. Asexual *P. falciparum* NF54 parasite cultures were maintained at 5-8% parasitaemia, 5% haematocrit in human erythrocytes at 37°C in RPMI 1640 medium supplemented as described under hypoxic

conditions (90% N_2, 5% O_2, and 5% CO_2)[110,111] with shaking at 60 rpm. Synchronous asexual cultures (>95% synchronicity of ring-stage parasites, 3 hr window) was obtained by at least three consecutive cycles of treatment with 5% D-sorbitol, each 6-8 h apart. All cultures were maintained with daily medium changes and monitored with Giemsa-stained thin smear microscopy.

Arrest was induced by removing the polyamine putrescine as key mitogen required for parasite cell cycle progression, by the addition of a specific inhibitor, difluoromethylornithine (DMFO). This caused an arrest at the G1/S transition point of the parasite's cell cycle. The arrest was then reversed by the addition of exogenous putrescine, which allowed the parasite to re-enter (RE) the cell cycle, with samples taken at consecutive timepoints of 3 h (RE1), 6 h (RE2) and 12 h (RE3) following reversal (50 ml samples at 5% haematocrit, 10-15% parasitaemia). cDNA microarray was performed on custom microarray slides that contained 12 468 oligos (60-mer, Agilent Technologies). Differentially expressed (DE) genes were identified after robust-spline within-slide normalization and Gquantile between-slide normalization as genes with $\log_2$ fold change ($\log_2$ FC of untreated [UT] / treated [T]) of 0.75 in either increased or decreased abundance. All experiments were performed by Dr. R. van Biljon[112]. The DE gene sets from the respective arrested transcriptome and the RE points where subsequently used downstream to inform regulatory candidates in the GRN construction.

2.2.1.2 GRN construction from asexual proliferation control points

A Gene Association Network (GAN) for *P. falciparum* from STRING v 10.0 [113] (https://string-db.org/) was filtered for DE genes present in the datasets for the RE1 and RE2 time points. STRING categories used consisted of experimentally determined interactions, curated database references, fusion data as well as co-expression. A combined score was calculated for these categories in accordance with STRING guidelines and a threshold of ≥ 0.8 was imposed on the data.

Differentially expressed ($\log_2$ FC>0.75) in the RE1 and RE2 timepoints were used to inform the construction of a GRENITS-based GRN. The postulation was that transcripts that were immediately differentially expressed upon re-entry into the cell cycle should have importance and drive subsequent processes required for the continuation of the cell cycle and ultimately result in proliferation. These putative "regulators" were subsequently evaluated through the GRENITS package in R [109] with an oligonucleotide DNA microarray dataset, which captured hourly transcription during asexual development over the 48 h cycle (GSE66669) [48]. This dataset provides the most comprehensive time series dataset of asexual development clearly

underpinning the progression of transcripts in the IDC. GRENITS constructs a linear interaction model through Gibbs variable selection on a DBN. Differentially expressed genes from the RE1 and RE2 timepoints were used as regulators and evaluated target genes from the RE2 and RE3 analysis. A probability threshold > 0.25 was considered for the interactions. Consolidations of the GRENITS network and the GAN was done based on edge reference.

2.2.2 Gene regulatory network of *P. falciparum* sexual differentiation

2.2.2.1 Parasite culturing and sampling

Asexual *P. falciparum* NF54 parasite cultures (NF54-*pfs16*-GFP-Luc,[114]) were maintained at 5-8% parasitemia 37°C in human erythrocytes at 5% haematocrit in RPMI 1640 medium supplemented as before under hypoxic conditions [110,111]. Gametocytogenesis was induced from 3x synchronized asexual parasite cultures by employing a strategy of concurrent nutrient starvation and a decrease of hematocrit[115]. Ring-stage parasite cultures (>95%) were adjusted to a 0.5% parasitaemia, 6% haematocrit in glucose-free medium containing culture (day -3) and maintained under the same hypoxic conditions at 37°C without shaking. After 72 h, the haematocrit was adjusted to 3% (day 0) and after another 24 h, the glucose-free medium was replaced with medium containing 0.2% D-glucose and this maintained for the duration of gametocytogenesis for a total of 16 days. Contaminating asexual parasites were removed daily with 5% D-sorbitol and/or *N*-acetylglucosamine treatment for 15 minutes at 37°C.

Samples (30 ml of 2-3% gametocytaemia, 4-6% haematocrit) were taken daily for microarray analysis on days -2 to 13 following gametocyte induction and enriched for gametocytes with either 0.01% w/v saponin treatment (3 min, 22°C) for early-stage gametocytes (stage I-III) or via density centrifugation (20 min at 800x*g)* using Nycoprep 1.077 cushions (Axis-Shield) for late stage gametocyte (stage IV/V) [115].

DNA microarrays were performed on custom Agilent 60-mer 8x15k arrays (AMADID#037237) and scanned on an Agilent G2600D Microarray Scanner (Agilent Technologies, USA) data were included that had both red and green intensities that were well above background and passed spot filters (P<0.01) and $\log_2$ (Cy5/Cy3) expression values. Genes that varied significantly during gametocytogenesis were identified with a one class, two-tailed t-test with an alpha value of 0.05 using Welch approximation and testing against mean=0. Data were clustered hierarchically according to Euclidean distance with average linkage clustering. All experiments were performed by Dr. R. van Biljon[112] (GSE104889).

Genes that varied significantly during gametocytogenesis were identified by conducting a one class, two-tailed t-test with an alpha value of 0.05 using welch approximation and testing against mean=0. Data were clustered hierarchically according to Euclidean distance with average linkage clustering. Pearson correlation coefficients were calculated in the R statistical package (version 3.2.3) comparing the expression of genes at each time point with every other time point and visualized using the *corrplot* package in R. Genes involved in different metabolic pathways were obtained from the Malaria Parasite Metabolic Pathways (MPMP) database and sourced from the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[116,117]. Subsets of genes related to functional categories were obtained from PlasmoDB, using search terms histone*, kinase, ApiAP2, signalling and signalling transduction along with published phosphatases[118] and epigenetic factors[119]. Significantly affected genes in these subsets were identified by ranking increased and decreased genes with an alpha value of 0.05 using the entire time course as input.

2.2.2.2 Gene association network for gametocytogenesis

Genes that varied significantly during gametocytogenesis were used to map a gene association network (GAN) from the STRING database v10.0 in R[113]. The GAN was used to show the association between genes based on co-expression, homology, protein-protein interactions, *in silico* predictions etc. and mapped with a combined confidence score greater than 0.7 (text-mining data excluded) to produce a high confidence network. Subsequently interacting nodes (interacting with the probe nodes) were included in the network mapping, with a total of 1350 nodes and 3357 edges. Sub-graphing of the original network was performed to highlight interactions of interest after significant gametocyte nodes with >50 interactors were decreased by removing interacting nodes that did not vary substantially throughout maturation for visualization in (746 nodes and 1616 edges). Network visualization was done using igraph package in R using the Fruchter-Reingold layout algorithm. Nodes were colour scaled according to expression values[120].

2.2.2.3 GRN construction for sexual differentiation

A total set of 13 ApiAP2 transcription factors (all ApiAP2's with increased abundance during gametocyte development) were used to evaluate their possible function in the regulation of sexual development across the 16-day time series. A GRN was constructed using GRENITS and using the 13 ApiAP2's as regulators with a probability threshold >0.7 against the microarray dataset (GSE104889). The number of links per model, per threshold was evaluated to determine the set probability threshold for the constructed GRENITS network (probability

threshold >0.7). Visualization of the GRENITS network was performed using the igraph package in R[120].

2.3 Results:

2.3.1 Candidates for cell cycle regulation in *P. falciparum* asexual proliferation inferred as a result of strategic arrest and re-entry

Strategic cell cycle arrest and re-entry at the G1/S transition point within asexual proliferation of *P. falciparum* parasites[96] (Figure 2.1), resulted in a dataset where candidate genes with the potential to drive developmental regulation during synchronized cell cycle re-entry can be inferred. The arrest was based on the treatment of asexual parasites during its cell cycle G1 phase (ring-stage parasites at 0-12 h post-invasion) with DFMO as suicide inhibitor of ornithine decarboxylase. This prevents the synthesis of polyamines as mitogens for cell cycle continuation and arrests parasite proliferation in early trophozoite stages, corresponding to the parasite's G1/S transition point of its cell cycle. This arrest can then be reversed with addition of the mitogen (putrescine). The cell cycle re-entry was monitored over 12 h with samples taken at 3, 6 and 12 h after re-entry (RE1-3, Figure 2.1).

This system therefore allowed sampling and evaluation of the asexual parasite's transcriptome at discrete points corresponding to the cell cycle to determine regulatory elements. As such, DNA microarrays were performed on cell cycle arrested parasites and this transcriptome compared to the differential expression ($\log_2$ FC >0.75 increased or decreased, FC=UT/T) of transcripts during the three re-entry points (Figure 2.2).

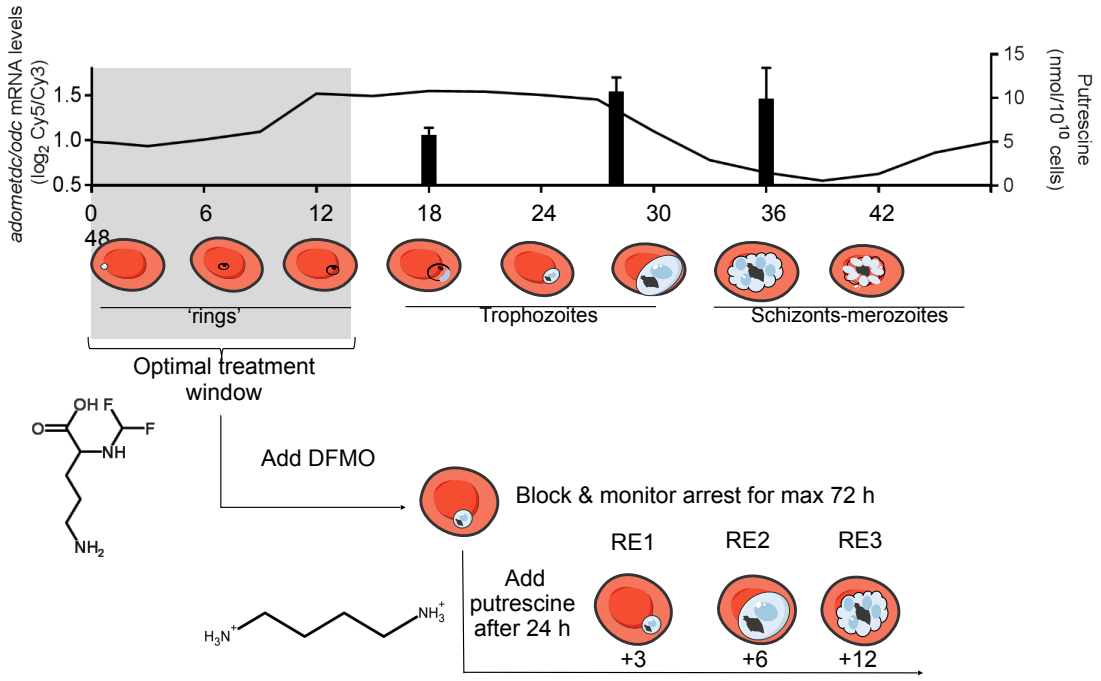


Figure 2.1: *P. falciparum* cell cycle regulation at the G1/S transition point.
Cell cycle arrest was induced in asexual *P. falciparum* parasite by treatment with difluoromethyl ornithine (DFMO) as suicide inhibitor of PfAdoMetDC/ODC, which is expressed from 12 h post-infection (hpi). This results in parasite cell cycle arrest in early trophozoite stages (corresponding to parasites at 18-22 hpi). The arrest can be reversed by the addition of putrescine as mitogen, resulting in parasites re-initiating cell cycle development. Samples taken at various timepoints after re-initiation provide data on processes involved in cell cycle control. RE1: re-entry timepoint 1, RE2: re-entry timepoint 2 and RE3: re-entry timepoint 3. [94].

The cell cycle arrest at the G_1/S transition point, showed 988 differentially increased transcripts, while 975 decreased (Figure 2.2 A), showing that about a third of the parasite's genome is affected between the point of arrest and that these changes are reversed after progression through the cell cycle has been re-established. These differentially perturbed transcripts were further investigated for 4 functional clusters of genes: (i) kinases and phosphatases (228 transcripts), (ii) DNA replication (73 transcripts), (iii) transcription and chromatin dynamics (272 transcripts) and (iv) Ca^{2+} signalling associated processes (95 transcripts). The dynamics of the DE genes was assessed by probing for their occurrence either immediately upon re-entry in RE1 as early responders (14 genes) or subsequently during RE2 (intermediate responders) or RE3 (late responders) (Figure 2.2 B). Re-entry into the cell cycle was characterised by the immediate expression of *myb1* and *cdpk4* upon polyamine supply, whilst *sap18* and *pk2* both display decreased expression upon cell cycle re-entry. After this initial expression of genes, a second subset are associated with RE2 (n=91), including kinases (*pk5*, *pkac* and *pkar), set9*, ApiAP2 *pf3d7_1115500,* DNA primase and an unannotated gene. The latter, *pf3d7_0105800,* contained an Interpro domain associated with cyclin dependent kinases (IPR000789). Genes involved in the late response

(RE3) are enriched in transcription factors (3 ApiAP2 family members) and CDKs and Aurora kinases (*crk3, 4, 5* and *ark2*), with increased expression of pre-RC genes *orc1&2* and *mcm3, 4, 5, 6, 8* also evident.

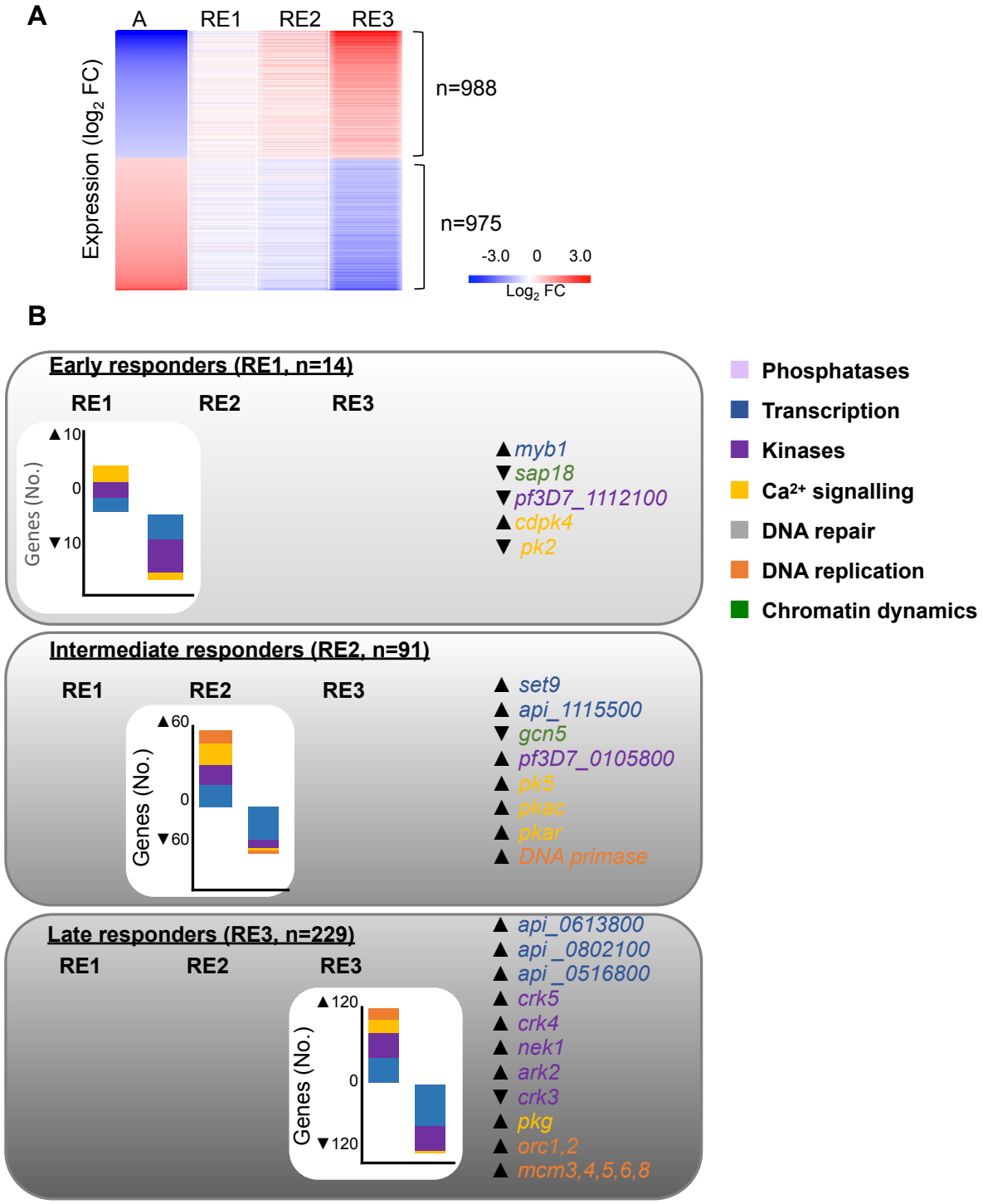


Figure 2.2: Molecular mechanisms controlling cell cycle re-entry.
A) The transcriptomes of parasites that re-entered their cell cycles (RE1-3) were analysed in context of genes matching key terms associated with cell cycle regulation using PlasmoDB (v33). B) The number of DE (either increased or decreased abundance) genes associated with specific cell cycle-related functional associations in histograms and genes of interest highlighted in grey boxes.

Gene names: *myb1: pf3d7_1315800* transcription factor MYB1, *sap18: pf3d7_0711400* histone deacetylase complex subunit SAP18, *cdpk4: pf3d7_0717500* calcium-dependent protein kinase 4, *pk2: pf3d7_1238900* protein kinase 2, *set9: pf3d7_0508100* SET domain protein, putative, *gcn5: pf3d7_0823300* histone acetyltransferase GCN5, *pk5*: *pf3d7_1356900* protein kinase 5, *pkac: pf3d7_0934800* cAMP-dependent protein kinase catalytic subunit, *pkar: pf3d7_1223100* cAMP-dependent protein kinase regulatory subunit, *crk5: pf3d7_0615500* cdc2-related protein kinase 5, *crk4: pf3d7_0317200* cdc2-related protein kinase 4, *nek1: pf3d7_1228300* NIMA related kinase 1, *ark2: pf3d7_0309200* serine/threonine protein kinase ARK2, putative, *crk3: pf3d7_0415300* cdc2-related protein kinase 3, *pkg: pf3d7_1436600* cGMP-dependent protein kinase, *orc:* origin recognition complex subunit (*pf3d7_1203000, pf3d7_0705300)* and *mcm:* DNA replication licensing factor (*pf3d7_0527000, pf3d7_1317100, pf3d7_1211700, pf3d7_1355100, pf3d7_1211300*).

The identification of the early-responder genes in RE1 enable inference of directionality of gene regulation for our putative regulatory candidates: ie RE1→RE2→RE3; RE1→RE3 or RE2→RE3. We used the dynamics in expression of cell cycle-associated genes over the full 48 h cycle[49] to predict directional relationships between particularly important gene nodes during re-entry with the GRN, based on co-expression data obtained at various re-entry time points. These regulatory candidates and their resolved target genes were used then to construct an interconnected subnetwork of factors driving the cell cycle progression (Figure 2.3). GRENITS probabilities >0.25 and STRING combined scores >0.8 was used to define the edges of the subnetwork.

The subnetwork revealed 6 key nodes that were DE in RE1 and started the initial cascade of gene expression as the parasite progresses through its cell cycle (Figure 2.3). In the 1st arm, Ca^{2+} signalling was shown to play a primary role in enabling cell cycle progression. The transcripts for *cdpk4* and *pk2* were closely interconnected nodes, both acting on downstream kinases and Ca^{2+} signalling machinery (e.g. *pkg, pkac* and *pkar*). Interestingly, *cdpk4* increased in transcript abundance while *pk2* decreased and showed a directional interaction from *cdpk4* to *pk2*, suggesting that *cdpk4* might regulate *pk2*, decreasing its expression for the cell cycle to progress. *Cdpk4* also had 58 direct connections to transcripts involved in Ca^{2+} signalling, transcription, kinases and phosphatases, making it one of the central nodes in this network (Figure 2.3). This suggests that CDPK4 expression could stimulate entry into S-phase and lead to the expression of genes essential for completion of schizogony in *P. falciparum,* similar to its requirement for entry into S-phase in *P. berghei* gametogenesis[18].

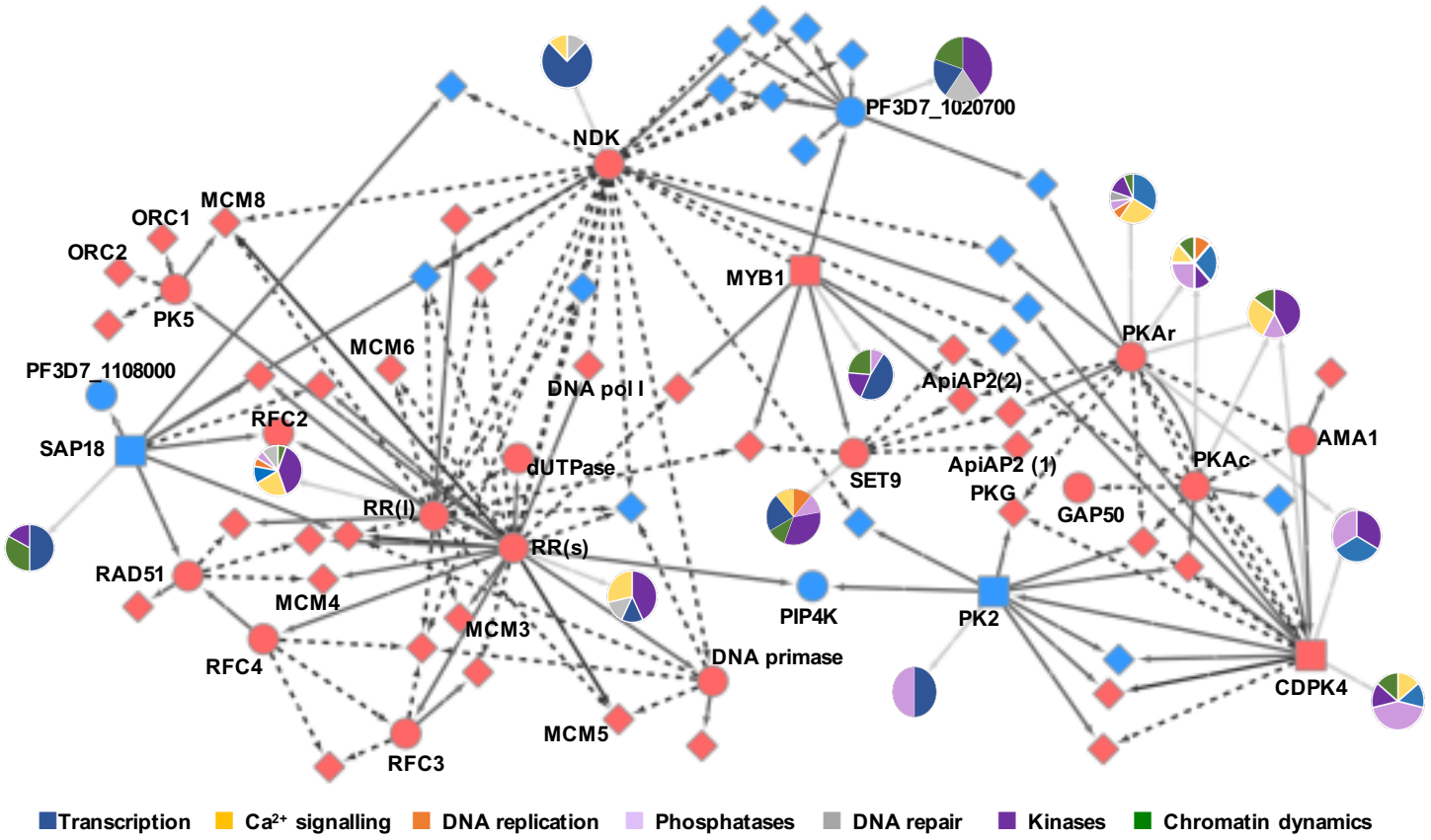


Figure 2.3: A GRN generated by GRENITS on differentially expressed genes driving cell cycle re-entry in asexual *P. falciparum* parasites.
A GRN was constructed between putative regulators of cell cycle re-entry by combining co-expression analysis (GRENITS, probability linkage score > 0.25) shown as solid grey edges and functional association between genes (STRING v 10.0, combined probability score > 0.8) shown as dashed black edges, while associations stemming from both analyses are shown as solid black lines. Transcripts that were differentially expressed by RE1 (■), RE2 (●) and RE3 (♦) are indicated according to their increased (red) or decreased (blue) transcript abundance. Genes of interest are indicated by gene symbol on the network, (RR = ribonucleoside-diphosphate reductase large (l) subunit, *pf3d7_1437200* and small subunit (s), *pf3d7_1015800*, ApiAP2 (1) = *pf3d7_0613800*, ApiAP2 (2) = *pf3d7_0802100*) while genes that were not subsequently analysed were collapsed into pie diagrams in terms of their functional cluster in the legend. Gene names: *myb1: pf3d7_1315800* transcription factor MYB1, *sap18: pf3d7_0711400* histone deacetylase complex subunit SAP18, *cdpk4: pf3d7_0717500* calcium-dependent protein kinase 4, *pk2: pf3d7_1238900* protein kinase 2, *set9: pf3d7_0508100* SET domain protein, putative, *gcn5: pf3d7_0823300* histone acetyltransferase GCN5, *pk5*: *pf3d7_1356900* protein kinase 5, *pkac: pf3d7_0934800* cAMP-dependent protein kinase catalytic subunit, *pkar: pf3d7_1223100* cAMP-dependent protein kinase regulatory subunit, *crk5: pf3d7_0615500* cdc2-related protein kinase 5, *crk4: pf3d7_0317200* cdc2-related protein kinase 4, *nek1: pf3d7_1228300* NIMA related kinase 1, *ark2: pf3d7_0309200* serine/threonine protein kinase ARK2, putative, *crk3: pf3d7_0415300* cdc2-related protein kinase 3, *pkg: pf3d7_1436600* cGMP-dependent protein kinase, *orc:* origin recognition complex subunit (*pf3d7_1203000, pf3d7_0705300), mcm:* DNA replication licensing factor (*pf3d7_0527000, pf3d7_1317100, pf3d7_1211700, pf3d7_1355100, pf3d7_1211300*), *pf3d7_1133400* apical membrane antigen 1, *pf3d7_0918000* glideosome-associated protein 50, *pf3d7_1366500* nucleoside diphosphate kinase and *pf3d7_1107400* DNA repair protein RAD51.

Importantly, the transcription factor *myb1* is indicated as a differential transcript in RE1[47]. The potential importance of *myb1* as a regulator of cell cycle control is emphasized by it connecting, and per implication regulating, 12 other cell-cycle related genes. Of significance is the fact that *myb1* connects with *cdpk4* and two ApiAP2 transcription factors (*pf3d7_0802100* and *pf3d7_0613800*, also DE in RE3). This implies that *myb1* could be involved in sequence-specific transcriptional regulation of other transcription factors and genes involved in signalling events. MYB1 also directly associates with epigenetic factors such putative histone acetyl transferase *set9* (in the RE2 set) and *pf3d7_1020700* (decreased in RE2). The GRN therefore implicates *myb1* as an important and central regulator of gene expression during cell cycle regulation of malaria parasites.

The possible role of epigenetic mechanisms involved in cell cycle regulation extends further through decreased abundance of *sap18* (RE1), which potentially permits the expression of pre-replicative licensing factors (*mcm3-6*), DNA polymerases and other replication factors (including *rad51, rec2* and *pf3d7_1108000*) as indicated by the interactions between *sap18* as controlling node to these transcripts. Furthermore, *pf3d7_1020700*, a putative histone acetyltransferase, is also highly connected (nine connections) and the genes that it connects to are all involved in transcriptional events. Both *sap18* and *pf3d7_1020700* have not been functionally characterized to any extent in *P. falciparum*. This highlights the power and importance of GRNs, such that genes are ascribed with potential important regulatory functions to control cell cycle regulation in asexual *P. falciparum* parasites, that would not otherwise have been identified.

Taken together, the GRN in this instance was applied as a powerful tool to identify novel regulatory elements that control proliferation of malaria parasites and elucidate the molecular determinants of a cell cycle checkpoint in the normally asynchronously dividing asexual parasite.

2.3.2 The molecular landscape undergoes specific changes as gametocytes progress through development

Gene-association network analysis (STRING) was used to associate functionality to transcripts that were significantly increased in abundance during gametocyte development. A complete transcriptome was generated for each day of gametocyte development and provided a high-resolution dataset to explore for regulatory elements. Unlike for the cell cycle, the question of regulation in this instance was much broader, making it difficult to apply a dynamic

Bayesian network in the same way, as there were not putative regulators that could be used as a starting point for this study. Thus, here we applied a more fundamental network approach to provide an overview of how gene expression changes throughout gametocytogenesis, before building out more complicated approaches in the rest of the thesis. Transcripts with significantly increased abundance during gametocytogenesis (α<0.05) were mapped to their interacting partners to identify processes that are regulated during specific times in gametocyte development (Figure 2.4). Two clusters of transcripts were increased in abundance immediately at the onset of gametocyte development and characterized stage I (day 3) gametocytes. Within these, biological enrichment for transcripts associated with fatty acid biosynthesis and biotin metabolism (GO: fatty acid biosynthesis, fold enrichment=29.41, P=1.91x10^{-8}; GO: biotin metabolism, fold enrichment=33.62, P=2.19x10^{-5}) was observed. The cluster enriched for fatty acid and biotin metabolism included five genes that showed significantly increased transcript abundance during gametocyte development and include enoyl-acyl carrier reductase (*enr, pf3d7_0615100*), malonyl CoA-acyl carrier protein transacylase precursor (*mcat, pf3d7_1312000*), 3-oxoacyl-[acyl-carrier-protein] reductase (*FabG, pf3d7_0922900*), beta-ketoacyl-ACP synthase III (*kasIII, pf3d7_0211400*) and the biotin carboxylase subunit of acetyl CoA carboxylase (putative ACC, *pf3d7_1469600*). These genes interact with genes also involved in fatty acid storage and metabolism as well as the tricarboxylic acid (TCA) cycle. One of the interactors, acyl-CoA synthetase 9 (*pf3d7_0215000*), is expressed throughout gametocytogenesis. Interestingly, while seven of the acyl-CoA synthetase genes show increased expression during asexual development[121], only the parent gene[122] (acyl-CoA synthetase 9, *pf3d7_0215000*) is expressed during gametocytogenesis. These results indicate that the differences between asexual and gametocyte metabolism may involve more processes than previously described[14,123].

Transition from the onset of gametocyte development (stage I) to gametocyte differentiation (stage IIa) is characterized by the significant increase in abundance in transcripts associated in a cluster involved in microtubule-based movement, cellular component movement and microtubule-based processes, from day 4 onwards (GO: protein polymerization, fold enrichment=86.82, P=2.72x10^{-8}, GO: mitochondrial respiration, fold enrichment=37.21, P=6.48x10^{-7} GO: nucleotide metabolism fold enrichment=32.9, P=8.03x10^{-8}). The increase in expression of genes involved in the TCA cycle and mitochondrial respiration correlates with metabolomic studies of gametocytogenesis[14,123] and were characterized by a concurrent decrease of expression in genes involved in glycolysis. Several of the transcripts (e.g. *pf3d7_1020100, pf3d7_1122900, pf3d7_0906400, pf3d7_1020300*) and their interactors form cytoskeletal components and include a microtubule binding protein EB1 homologue (*pf3d7_0307300*), which interact with *ark1, ark2* and *ark3* (*pf3d7_0605300, pf3d7_0309200,*

pf3d7_1356800). These kinases show differential expression throughout gametocytogenesis: *ark2* with increased transcript abundance throughout development; *ark1* is increased from stage II-IV and *ark3* only increased during stage IV-V development. Within this cluster, adenylate kinase 2 (*pf3d7_0816900*) interacts with 14 genes mostly involved in mitochondrial respiration and nucleotide metabolism. Overall, this cluster highlights many known aspects of the divergence from asexual to sexual parasites, namely an increased reliance on mitochondrial metabolism and the cytoskeletal remodelling gametocytes undergo to allow for their uniquely elongated shape[14,124].

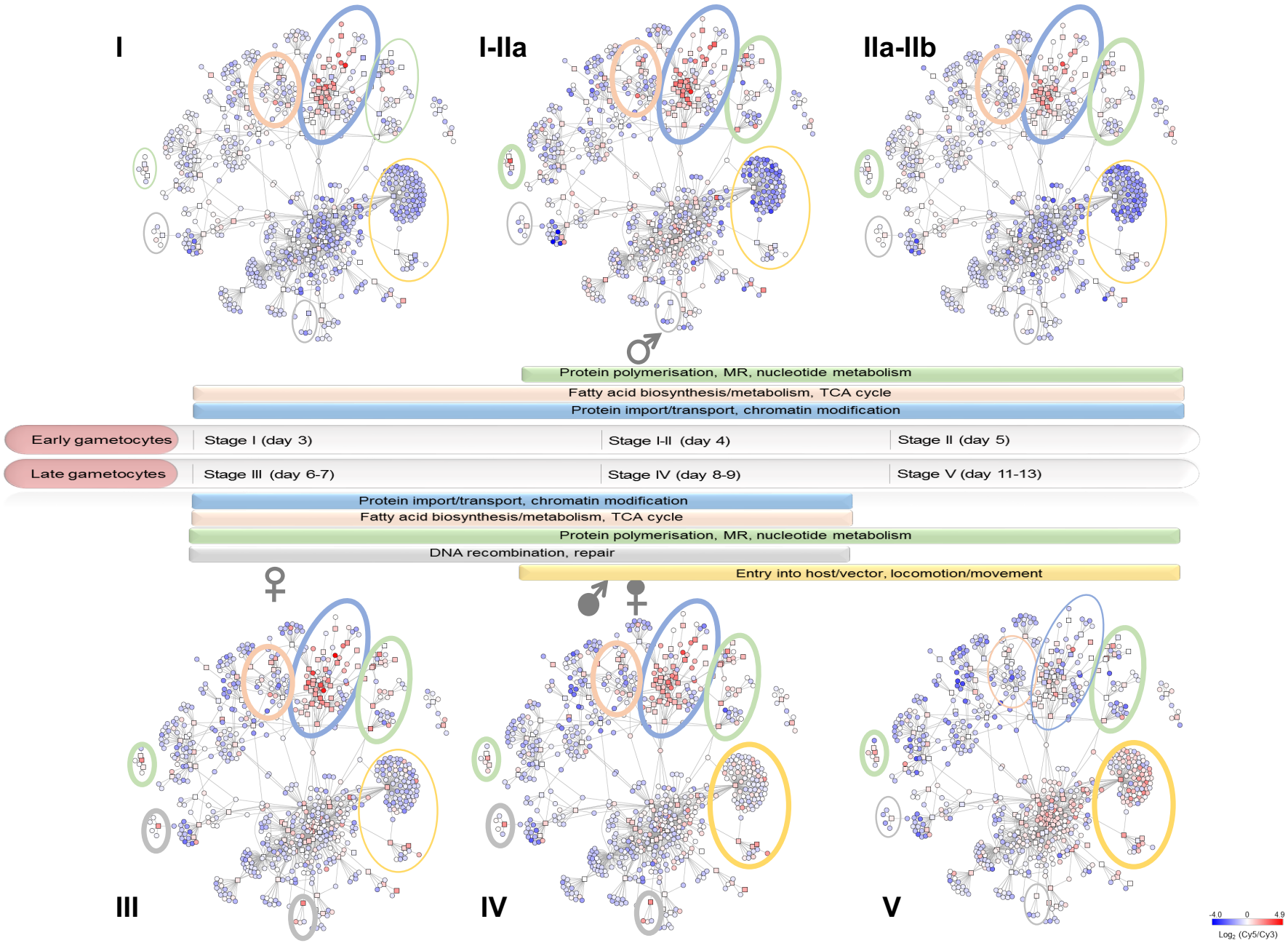

I
I-IIa
IIa-IIb
Protein polymerisation, MR, nucleotide metabolism
Fatty acid biosynthesis/metabolism, TCA cycle
Protein import/transport, chromatin modification
Early gametocytes
Stage I (day 3)
Stage I-II (day 4)
Stage II (day 5)
Late gametocytes
Stage III (day 6-7)
Stage IV (day 8-9)
Stage V (day 11-13)
Protein import/transport, chromatin modification
Fatty acid biosynthesis/metabolism, TCA cycle
Protein polymerisation, MR, nucleotide metabolism
DNA recombination, repair
Entry into host/vector, locomotion/movement
III
IV
V
-4.0
0
4.9
Log2 (Cy5/Cy3)

Figure 2.4: Gene interaction network analysis reveals the molecular landscape of gametocyte maturation.
Genes that were significantly increased in expression during gametocyte were mapped along with their subsequent interactions (STRING database network with combined confidence threshold >0.7) to identify key transcriptional role-players that shape the gametocyte transcriptome. Interactions were mapped to days corresponding to the developmental stages described below each network. For visual clarity, nodes with more than 50 interactions had interactors that do not vary substantially in expression omitted from the Figure. Coloured bars indicate a summary of enriched GO terms within clusters of genes expressing during each developmental stage with the corresponding clusters circled with matching colours. The weight of the circled lines correlate to the increased number of transcripts within the cluster that showed increased abundance. Genes that characterize sex-specific development were monitored over the time course and the start and peak of their increase in transcript abundance indicated for males and females with open symbols indicating start of expression and filled symbols indicating peak of expression.

Stage III-IV development is characterized by induction of two small clusters involved in DNA recombination and repair (GO: DNA recombination, fold enrichment=10.47, P=8.56x10^{-4}, GO: DNA repair, fold enrichment=7.18, P=2.41x10^{-5}). Two significantly expressed transcripts include meiotic recombination protein DMC1 (*pf3d7_0816800*) and RuvB-like helicase 1 (*ruvb1, pf3d7_0809700*). DMC1 interacted with DNA repair and transcripts for recombination proteins including *rad54* (*pf3d7_0803400*) and *rad51* (*pf3d7_1107400*) whilst *ruvb1*interacted with histone H2A variant H2A.Z (*pf3d7_0320900*) signifying the transition of the gametocyte in starting to prepare for gametogenesis in these stages.

A strong cluster with transcripts involved in protein import and localization into the nucleus and chromatin modification was present throughout stage I-IV development (days 3-10) (GO: protein import into nucleus, fold enrichment=23.21, P= 8.96x10^{-5}; and GO: chromatin modification, fold enrichment=10.09, P=9.96x10^{-4}). This gene cluster was unique as 23 out of 32 significantly expressed transcripts in this cluster were highly increased in abundance ($\log_2$ (Cy5/Cy3) of >0.5) and these transcripts were also the most highly connected to each other; 26 of the 32 transcripts interacted with at least one other significantly regulated gene. For instance, the cluster contains the CCR4-NOT transcription complex subunit 2 (*not2*, *pf3d7_1128600*) which is significantly expressed (P<0.05), associated with interacting partners including other members of the CCR4-NOT transcription complex, of which all but CCR4-associated factor 1 *caf1* (*pf3d7_0811300*) and a NOT family protein (*pf3d7_1417200*) are increased in abundance during gametocyte development.

The final stages of gametocyte maturation (stage IV-V differentiation) were characterized by the absence of all clusters associated with earlier stages of development (except for protein polymerization, mitochondrial respiration, and nucleotide metabolism) and the presence of a cluster of transcripts associated with host/vector entry and locomotion (GO: entry into host or other organism, fold enrichment=9, P=2.42x10^{-5}, and GO: locomotion fold enrichment=6.96,

P=1.05x10^{-4}). The cluster contained four genes significantly expressed, with *pf3d7_0914100* (unknown function) the most associated gene, interacting with 125 other genes (96 interactions displayed in Figure 2.4). These interactors include *ark3*, *cdpk4* (*pf3d7_0717500*), the ookinete-associated transcription factor *ap2-o* (*pf3d7_1143100*), dynein light chain (*pf3d7_0729800*) and a protein phosphatase (*ppm5, pf3d7_0810300*) as well as a number of genes involved in invasion including reticulocyte binding homologues *Rh1, 2a, 2b, 5*; erythrocyte binding antigen *eba140* (*pf3d7_1301600*), merozoite surface protein 11 (*msp11* (*pf3d7_1036000*). Several of these transcripts have been implicated in significant roles for completion of the mosquito stages of the life cycle (*ap2-O, cdpk4*[18]) and suggests that exported proteins that typically enable antigenic interactions in the blood stage has a further role in mosquito-stage development.

2.3.3 Transcriptional dynamics of sex-specificity and translationally repressed genes

Gametocytogenesis is also associated with the formation of male and female gametocytes. Transcripts associated with sex-specificity[22] were probed in the gametocyte transcriptome data and showed differential expression patterns. We show that sex-specification is a process initiated early during gametocyte development. Male-specific genes (n=528)[22] including male gamete fusion factor *hap2* (*pf3d7_1014200*), alpha tubulin 2 (*pf3d7_0422300*) and *cdpk4* (*pf3d7_0717500*) on average first start appearing on day 4 of development in stage IIa (Figure 2.4). The male specific genes gradually increase in expression and peak on day 9 of development in stage IV gametocytes ($\log_2$ (Cy5/Cy3) =0.45). Comparatively, female-specific genes (n=735)[22], including *pfs25* (*pf3d7_1031000*), ookinete surface protein *p28* (*pf3d7_1030900*) and transcription factor *ap2-O* (*pf3d7_1143100*) measurably increase in expression only from stage III (day 5/6) of development. However, peak expression of female-specific genes occurs slightly earlier than male-specific genes on day 8 ($\log_2$ (Cy5/Cy3) =0.74).

Translational repression of genes transcribed during gametocytogenesis is well documented in both *P. falciparum* and *P. berghei* gametocytes, particularly for mature stage V gametocytes [22,125–127]. However, the transcriptional dynamics of these transcripts during gametocytogenesis has not been investigated. We used the full gametocytogenesis transcriptome and mapped the putative translationally repressed genes (n=509)[22] thereto (Figure 2.4). Most of the translationally repressed genes (n=322) were not actively transcribed, however, two clusters of genes showed a concerted increase in expression throughout gametocytogenesis, the first of which already showed increased expression levels from stage I of development (n=90). The second cluster of genes increased in expression only from stage III of development and

included several ookinete associated proteins including *psop6, 13 and 20* (*pf3d7_0630200, pf3d7_0518800, pf3d7_0715400*), *ccp4* (*pf3d7_0903800*) and *p28* (*pf3d7_1030900*). This indicates that gametocytes already synthesized genes necessary for gametogenesis from the onset of sexual differentiation rather than just the late stages, which is the only point at which the transcript abundance of these genes were previously investigated.

2.3.4 Hierarchical contribution of transcription factors and kinases to gametocyte development

Given the broad involvement of ApiAP2 transcription factors and kinases to different life cycle stage progressions, individual transcripts from these two clusters were probed with GRENITS (Figure 2.5) to identify other associated genes in the gametocyte transcriptome that show patterns of expression that could result from direct or indirect regulation. The 13 important ApiAP2 family members as well as 40 kinases that were significantly increased in abundance were interrogated.

Amongst the kinases, 12 were highly connected (Figure 2.5) with possible regulatory activity, of which only *mapk1* (*pf3d7_1431500*) is putatively translationally repressed[22]. Along with *mapk1*, glycogen synthase kinase 3 (*gsk3*, *pf3d7_0312400*) and kinesin 13 (*klp8, pf3d7_1245100*) are associated with microtubule dynamics[128,129]. The NIMA-related kinase 4 (*nek-4, pf3d7_0719200)* was also associated with mitotic spindle microtubules[130] and has been shown to fulfil essential functions in the sexual stages of development[131], consistent with the gene product being involved in regulatory activity within the parasite. While *mapk1* and *klp8* are increased in abundance during the later stages of gametocytogenesis, *gsk3* is increased from stage I of development and *nek-4* is highly increased in expression between day 6-9 (stage III-IV) of development. The expression of these genes do taper down as the parasite enters stage V of development, as the rigid microtubule network present in the earlier stages of gametocyte development depolymerizes as the parasite matures[17,132]. This seems to suggest the involvement of kinases in the dynamics of the microtubule network constructed during sexual development[124,133].

Among the ApiAP2 transcription factors, five were shown to each regulate sets of ≥7 genes in the gametocyte transcriptome and these transcription factors show a cascade-like expression profile during gametocytogenesis (Figure 2.5). *Ap2-g* (*pf3d7_1222600*) is expressed first during gametocyte development and regulates a set of 19 genes[37]. *pf3d7_0934400* and *pf3d7_0611200* are successively expressed during the early stages of gametocyte

development, with each of these regulating sets of 12 and 36 genes, respectively. *pf3d7_0611200* interestingly also putatively regulates two other ApiAP2 transcription factors, *pf3d7_1107800* and *pf3d7_1139300*. As gametocytes develop into stage II-III, *pf3d7_1317200*, a ApiAP2 family member also previously associated with gametocyte regulation[37,106], peaks in expression and *pf3d7_1317200* regulates a smaller set of only seven genes. Although the putative regulated genes identified in the study for each of the transcription factors were queried for functional enrichment, none was found at $P<0.001$. In addition to the *in silico* analysis correlating the cascade-like expression profiles of the ApiAP2 members to possible regulatory targets, we also correlated the expression profiles of their experimentally validated target genes[37,106] as a direct indicator of functionality of these transcription factors in regulation of gametocytogenesis.

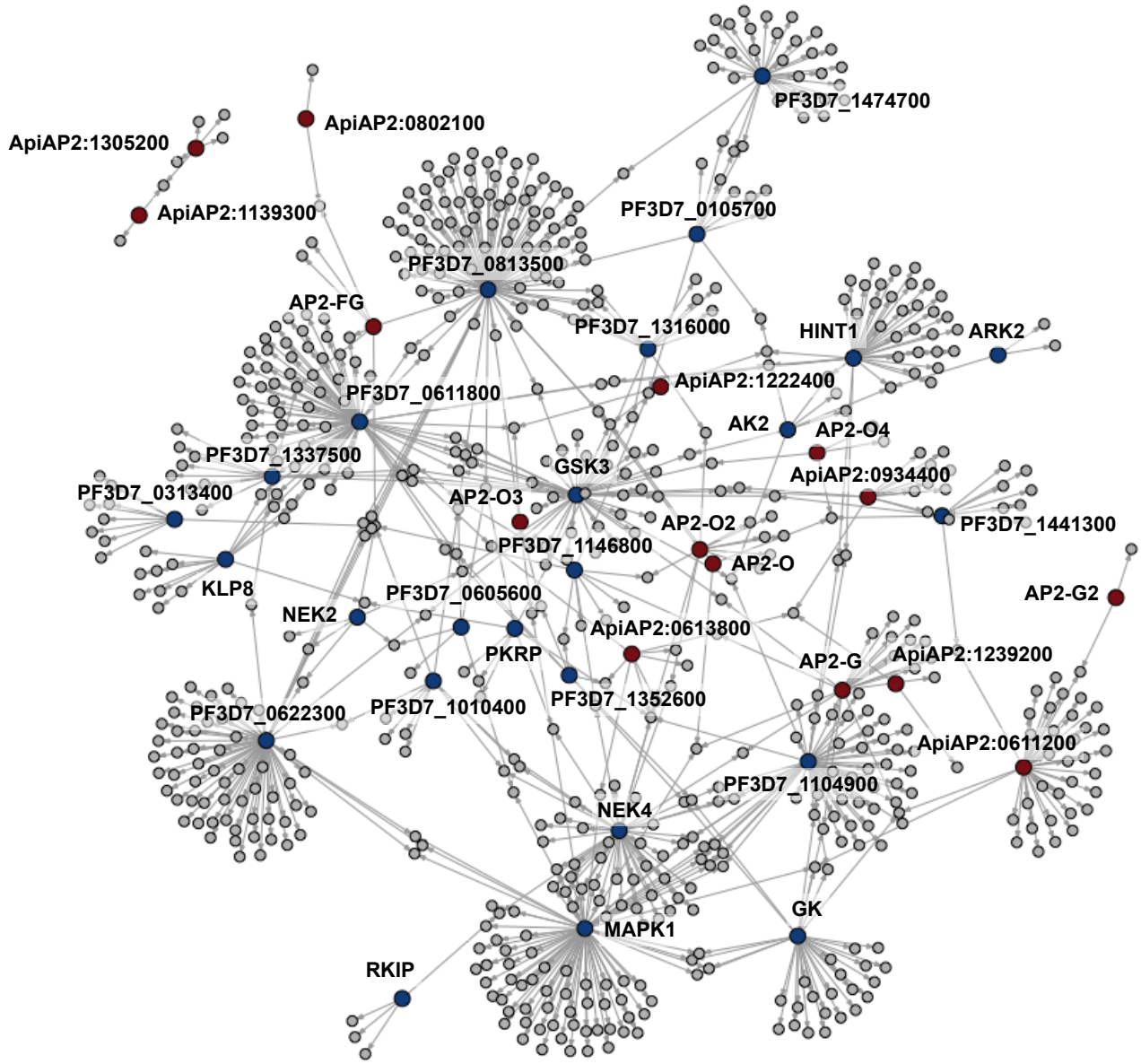


Figure 2.5: Specific regulatory elements contribute towards gametocytogenesis.
Putative regulators were chosen from significantly expressed kinases (P<0.05) and ApiAP2 transcription factors that show positive expression during gametocyte development. A gene regulation inference network was calculated (GRENITS with probability cutoff of >0.7) for the chosen ApiAP2 transcription factors (red nodes) and kinases (blue nodes) against the total gametocyte transcriptome to identify regulated nodes (grey nodes). Highly connected nodes are indicated with the corresponding PlasmoDB ID. Where available, annotated gene names are given: *GSK3; pf3d7_0312400, KLP8; pf3d7_1245100, AP2-G; pf3d7_1222600, GK; pf3d7_1351600, MAPK1; pf3d7_1431500, Nek-4; pf3d7_0719200*.

2.3.5 Resolution of ApiAP2 transcription factors driving gametocyte maturation

In order to further investigate the contribution of transcription factors associated with transcriptional regulation in gametocytogenesis, the ApiAP2 family of transcription factors

were closely interrogated for expression during the time course. Of the 27 ApiAP2's, only 15 were increased during gametocyte development. *Ap2-g* shows expected expression during asexual stages of development (Figure 2.6). Many ApiAP2's showed discrete increases during stage I to III intervals such as *pf3d7_0611200, pf3d7_0934400, pf3d7_1222400, pf3d7_1317200* and *pf3d7_1408200*, of which the latter two have been validated with knockout studies in rodent malaria[134]. Stably increased transcript levels for six ApiAP2's appear to be maintained during stage I to V (*pf3d7_0404100, pf3d7_0516800, pf3d7_0802100, pf3d7_1350900, pf3d7_1429200* and pf3d7_1449500. Late-stage gametocytes associated ApiAP2's showed *pf3d7_1143100, pf3d7_1239200* and *pf3d7_0613800* as increased during stage IV and V. *pf3d7_1143100* expression was shown to be translationally repressed in *P. berghei* gametocytes[134]. These transcription factors are therefore more likely to be functionally relevant during gamete stages in the parasite.

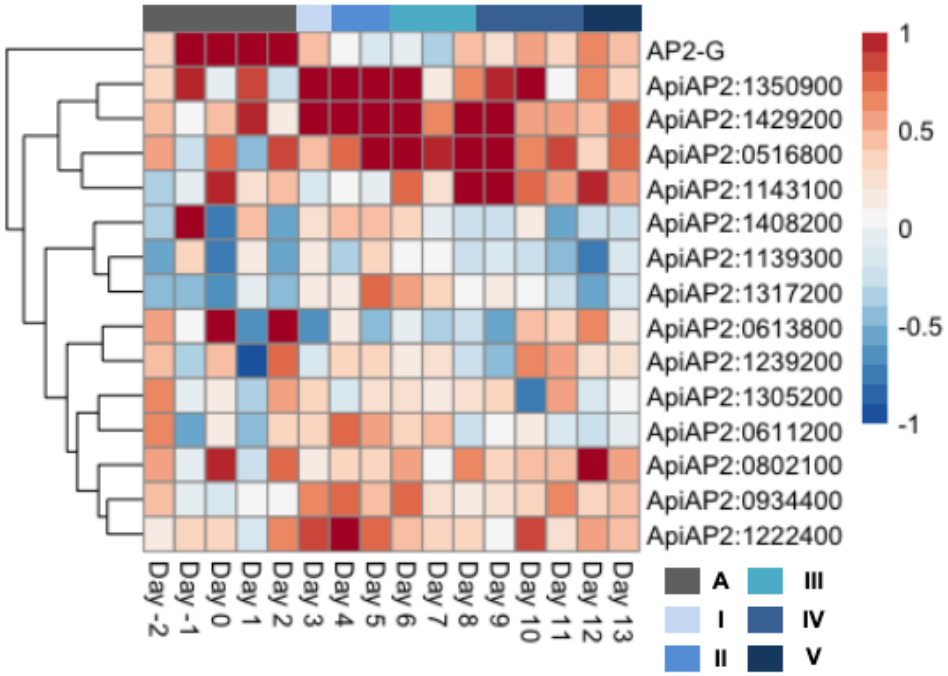


Figure 2.6: Expression of ApiAP2 during gametocyte development.
ApiAP2 transcription factors increased during gametocyte development. Gene ids are abbreviated as ApiAP2: followed by gene id. A = asexual stages, I = stage I, II = stage II, III = stage III, IV = stage IV and V = stage V.

Association of functional relevance is these ApiAP2's was done by incorporating GRENITS along with co-expressed gene sets. For *pf3d7_0611200,* a total of 314 genes co-expressed and 223 anti-correlated out of the proposed targets from the network (Figure 2.7). Gene ontology for anti-correlated targets shows these genes to mostly related to host invasion processes while co-expressed targets are implicated in RNA metabolism. Motif enrichment for 116 genes showed a TGCAC motif ($P = 5.5e^{-13}$, Figure 2.7) of which 100 were anti-correlated, indicating a repressive role. This motif is similar to the GTGCAC motif in AP2-I which could

suggest the role in the repression of invasion genes during gametocyte development by *pf3d7_0611200*. The second important ApiAP2 during gametocyte early development is *pf3d7_1317200*, with 21 targets in cell cycle processes such as DNA replication and chromosome organization. Unlike its ortholog in *P. berghei* (AP2-G3), no female specific enriched targets were found. Two ApiAP2's which are increased during stage I-V with *pf3d7_0934400* showing anti-correlated targets mostly (27/37 targets), indicating a possible need for a repressive role.

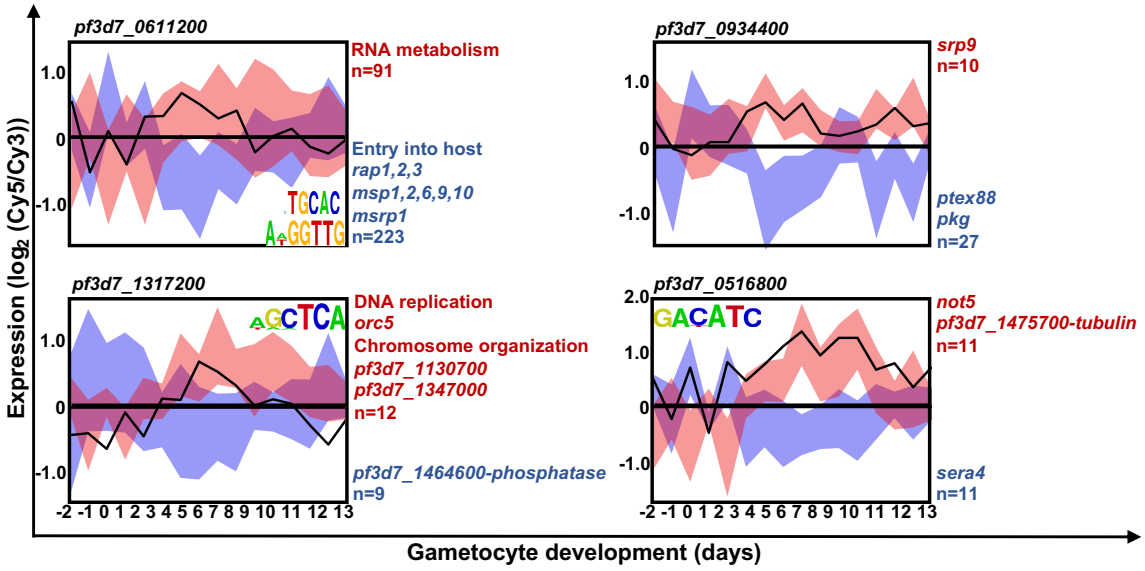


Figure 2.7: Specific regulatory elements contribute towards gametocytogenesis.
Highly connected ApiAP2 transcription factors that were identified as regulators were grouped by days on which they peak in transcript abundance (pink areas). Possible target genes were also highlighted from genes containing transcription factor binding sites (indicated on graphs) as well as from functional data indicating target genes for AP2-G.

2.4 Discussion:

The use of GRN analysis in *Plasmodium* research has been relatively scarce, while other fields have benefitted from this line of inferential reasoning. Many experimental investigations regarding regulatory mechanism and co-expression studies have been conducted, but the specific application of GRNs to these types of datasets are uncommon. Here, the study primarily focused on asexual proliferation and sexual differentiative development occurring in two different life phases of the parasite with the aim of determining candidates involved in regulating these developments. The global GRN, however, consolidates both these developmental phases while expanding on the number of evaluations conducted. This produced a large body of data in addition to the transcriptomes themselves and all aspects

these data cannot be discussed in detail but should rather be focused to select candidates of importance.

The use of dynamic Bayesian networks contributed to interesting findings for both IDC and gametocyte phases of development[94,95]. The role of Ca^{2+} signalling cascades with CDPK4, PKAc, PKAr, PK2 and PK5 were interconnected with transcription factors such as MYB1, ApiAP2's as well as epigenetic modifiers such as SET9 and SAP18. Constructing a targeted network through this approach using specific cell cycle re-entry points as a framework in conjunction with a high-resolution time course dataset, provided a unique and interesting method to investigate gene regulation in the parasite. This inference of directional associations, underpins the cell cycle cascade, showing the power and strength of GRNs in gene regulation research.

Differentiation during gametocyte development required a less directed approach, were inferring relevant biological features from established data sources such as STRING and a gametocyte time course produced interesting insights into transitions occurring during this phase. Given the more limited granularity of the gametocyte time course (fewer time points), the STRING network reduced the data to a simplified network highlighting key transitions which occur during gametocyte maturation. The associated morphological and physiological change in conversion to gametocytes occurs through activation of basic energy and macromolecule metabolism. Processes like glycolysis, the pentose phosphate pathway and hemoglobin catabolism are rapidly decreased as the gametocytes progress to stage II in development[105] and this is mimicked on the transcriptional level. Fatty acid biosynthesis seems to also be a hallmark of gametocyte development, with a number of genes involved in fatty acid biosynthesis in the apicoplast[135] being expressed throughout development. The increase of genes involved in the TCA cycle and oxidative phosphorylation also underline the importance of the increased mitochondrial cristae development in gametocytes[14,123,136].

Overall, this overview of the lifecycle highlighted a few key points in differentiation: the transition from stage I-II demarcates the preparation for the rapid transition to S-phase the parasite will undergo in the mosquito, with nucleotide metabolism and chromatin modification evident in stage IIa-IIb gametocytes. Subsequently, genes involved in chromatin modification and DNA recombination are more expressed in stage II-IV gametocytes. The early stages of gametocyte development additionally support sexual differentiation, with male sex-specific genes increasing in expression earlier (stage I-II) than female-specific genes (stage II-III)[22], corresponding to a slightly longer maturation time for female gametocytes[137]. From stage III onwards, the gametocytes are characterized by genes involved in transmission stages,

including those involved in meiotic DNA recombination and repair as well as entry into host and locomotion. The set of genes increased in expression during stage IV-V of development that are enriched for genes involved in entry into host and locomotion does support the transcription of genes involved in later stages of sexual development[22]. Finally, late stage V gametocytes ultimately have only a few essential metabolic processes activated[15,22,138], none of which are focused on energy metabolism. This results in a picture of a hypoactive stage V gametocyte, resulting in a reduced response to most drugs targeting metabolic processes and macromolecular synthesis[139,140]. This stage is characterized by a lull in transcriptional activity compared to the previous stages, with 1533 genes differentially expressed during this stage, the majority (802) being decreased in expression. The metabolic processes functionally enriched in gametocytes, fatty acid storage and the TCA cycle also typically characterize microbial quiescent cells[141]. Transcriptional quiescence regulators characterized in *Saccharomyces cerevisiae*, GSK3 and PKB[142] are increased in expression, while active cellular marker PKA is decreased in expression during gametocyte development.

Bayesian dynamic network analysis further resolved many of the ApiAP2s involved in driving gametocytogenesis. The expression of ApiAP2s was resolved previously for gametocyte development[105], however target genes for these transcription factors remains an open question. Many target genes for ApiAP2s were resolved as well as an ARID domain protein, which is considered a putative transcription factor. The transcriptional regulation of *P. falciparum* by ApiAP2s downstream of AP2-G shows a potential cascade of regulation during gametocytogenesis. The role of AP2-G2 and *pf3d7_0611200* in repressing transcription of asexual targets during gametocyte development are particularly interesting.

The use of GRNs were therefore important to define novel regulatory elements associated with the control of proliferation processes in *P. falciparum.* The insights produced through GRN construction proved a valuable resource, reducing the noise of data interpretation thorough prioritizing interaction sets throughout the time courses. Such patterns of probable causal relationships would have been hard to discern through methods outside the GRN space. Indeed, clustering and other approaches to evaluate the data as unsupervised methods are useful only to define general patterns for the data, but could not provide much information about the relationship between data points. GRNs produce quantified relationships between data points, which are more informative guides for understanding relationships in the data. These considerations were the key motivations behind using GRNs in parasite research for this chapter.

The fruitfulness of using GRNs really presented in how the network analysis were deployed. With a few well-plannedexperimental designs, such as the cell cycle arrest, we could evaluate a selection of interesting candidate genes based on their expression response following cell cycle re-entry and the known properties of the genes (kinases, transcription factors etc…). Often, the most complex part of the analysis is candidate selection. Transcription factors are the usual candidates of any GRN, but they are not the only genes which impact gene expression. Narrowing this list of candidates requires special consideration, such as the availability of known properties for the genes i.e. is there a biological basis for suspecting a regulatory role. Inclusion of irrelevant candidate genes (genes which might not be able to regulate or impact on gene expression) may introduce unnecessary noise and mislead the algorithms. Effectively this "boils down" to using proper feature selection and not constraining the model with erroneous artifacts which impede the analysis. This is where the arrest and re-entry experiment helped to guide the feature selection proses. We had fewer initiatives with the gametocyte candidates, thus we chiefly kept the GRENITS analysis to known transcription factors.

This strategy proved successful, developing two high quality networks for the different life phases of the parasite development. The probabilistic manner with which GRENITS resolved the relationships made the interpretations relatively intuitive and easy. Comparisons of the probability vs strength outputs as produced by GRENITS was also a necessary step to determine appropriate thresholding of subnetwork[109]. Once an observable divergence in the probability/strength relationship becomes clear, thresholding will be relatively easy to assign and meaningful probabilities can be filtered. Interpretations and inclusion of STRING data often required more considerations. It's not always clear from the STRING network how the data were derived. For example it's not always clear which transcriptomic data was used in the co-expression scores field or even more ambiguous the text-mining (association based) abstract scores. It's not clear how trusted some of the features are for the parasite biology, which is why we opted for inclusion of selected features and recalculated the combined scores (following their recommendations) to produce a more stringent network. This approach may have been overly stringent as we excluded a larger number of connections, but considering the unknown variable of the contributing data, we rather adopted a cautious approach in this case.

The insights obtained from this GRN based approached may have been difficult to elucidate otherwise, e.g. calcium signalling cascades with targets during the cell cycle progression and the role of repression for AP2-G2 during sexual development. The connections would have been hard to derive without GRNs. However, some difficulties where present for this approach

such as the scalability of DBNs[73,109]. We had to limit the analysis drastically towards essential comparisons, as the size of the proposed analysis would have been difficult to compute. This meant that we had to be more selective in our approach, which would imply many more missed connections which we could not obtain. The speed of the DBNs constructed here may in part be the result of the R programming language itself, given that the use of interpretated languages generally suffer memory and performance issues[143] and opting for or constructing a pipeline in C or other compiler based languages may have improved the computational efficiency as well as scalability. However, due to the extensive search function approaches required to define the underlying Bayesian networks (functions like MCMC) these algorithms are known for their excessive computation time[73]. This would always be a constraint in this approach. Even with these costly computations, we've shown strong features throughout the data and present a case for the use of GRNs in *P. falciparum* research.

We used limited resource networks such as GRENTIS and STRING in this chapter to guide questions regarding key regulators in specific cellular contexts. However, adapting these approaches to a wider, more encompassing, unsupervised approach such as a weighted co-expression network in chapter 3, allows expansion of the scope of investigations into the relevance of more potential regulatory candidate genes, particularly for gametocyte maturation.

Chapter 3

New insights into the sexual transcriptome of *Plasmodium falciparum* through high-resolution amplification-free RNA-seq and comprehensive gene regulatory network analysis

3.1 Introduction

The *Plasmodium* parasite completes its first phase of human infection in the liver before completing an indeterminate number of intraerythrocytic asexual developmental cycles. It is this 48 h asexual developmental cycle, where parasites mature from rings to trophozoites and finally schizonts to produce daughter merozoites, that gives rise to malaria symptoms and can lead to death in an untreated patient. However, to continue its life cycle beyond the human host, the parasite also forms sexual stages called gametocytes, the mature forms of which are the only stage that can transmit to the obligate *Anopheles* mosquito host. Sexual stage differentiation in the most lethal human malaria parasite, *Plasmodium falciparum*, is uniquely extended over a 10-14 day period[144] and is marked by development through five morphologically distinct developmental stages (stage I-V). This stage-specific transition is solely associated with this species of parasite, giving it its name based on the falciform shape of mature stage V gametocytes[15].

The ability of *P. falciparum* to transition between biologically distinct stages is mediated by tight control of gene expression. The transcriptome of both asexual parasites[20] as well as gametocytes[95] varies with every stage and most genes show conserved peak expression associated with specific life cycle stages[20,48,49,95,105]. This process is transcriptionally controlled by several mechanisms that are somewhat clarified for asexual development including the involvement of a small family of putative transcription factors[106], RNA decay mechanisms[48] and epigenetic mechanisms[28,33,45,48,145]. The functionally-probed family of transcription factors consist of only 27 members (ApiAP2 family, some of which are putative), which seems an unlikely small number of transcription factors to account for the regulation of all ~5500 genes known in the parasite[106]. An additional ~73 DNA-interacting proteins with likely transcription factor/associated domains have been identified primarily through *in silico* protein domain discovery and alignment strategies[146]. It is likely that these factors and even more may be involved in regulating transcripts. However, our understanding of processes driving transcriptional changes during stage transition associated with gametocytogenesis is limited to dynamic descriptors of the transcriptome from cDNA microarray data[95] and the chromatin

proteome[147] or to only the processes involved in the commitment steps changing asexual parasites into early stage gametocytes[33,44,45].

The application of RNA-seq technology has refined our understanding of gene models in the parasite[97], and identified the majority of splice sites including alternative splice sites. Extension of these investigations to single cell RNA-seq (scRNA-seq) yielded detailed insights into the early mechanisms employed by ring-stage asexual parasites to commit to sexual development[44,148]. The development of an unbiased RNA-seq protocol, directional amplification-free RNA-seq (DAFT-seq), has opened up avenues for the accurate evaluation of the *P. falciparum* transcriptome, even in light of the extraordinarily high AT-content of the genome (up to 95% in non-coding regions, compared to ~80% in gene bodies)[25]. Indeed, with this technology, the full power of RNA-seq over hybridisation-based transcriptomics is evident in the accurate description of non-coding RNAs (ncRNA) and long non-coding RNAs (lncRNA) as additional level of gene regulation used by the parasite[25]. Whilst previous RNA-seq reports[43,149–154] identified a large number of ncRNAs, the majority of these originate from AT-rich regions that were not accurately sequenced and assembled and as such have been discarded in reannotation processes[25]. The measure of known lncRNAs may be quantified with DNA-microarray, however due to the inherent nature of the platform, novel lncRNAs will be excluded. These lncRNAs can occur in two variants: intragenic and intergenic. For instance, intragenic lncRNA (often referred to as anti-sense RNA, asRNA) have been shown to suppress expression of an early driver of gametocytogenesis, gametocyte development protein 1 (GDV1)[45]. The mechanism by which asRNA silence gene expression in *Plasmodium* is still poorly understood as the parasite lacks dicer enzymes which are needed for RNA interference silencing mechanisms. Intergenic lncRNAs have also been implicated in gene regulation for the parasite such is the case with the telomere-associated long non-coding RNAs (lncRNA-TARE). As their names suggests these intergenic lncRNAs are situated near the telomeric regions of chromosomes and have been proposed to interfere with some members of the virulence-related *var* family of genes[155]. The proposed model for lncRNA-TARE is to interfere directly with *var* gene promoters or indirectly by recruiting histone modifying enzymes[155].

Here, we applied DAFT-seq to provide an in-depth evaluation of gene regulatory factors during the complete process of gametocyte development in *P. falciparum*. We produced a complete and high-resolution time course RNA-seq transcriptome for *P. falciparum* gametocyte development over a 16-day period, from commitment to gametocytogenesis through each stage of development and differentiation to the fully mature stage V gametocytes. Integration of this complete gametocyte dataset with existing gametocyte and asexual dataset produced

a large dataset spanning multiple stages of parasite development for comparison. We use these data to leverage the power of a phase-contrasting development model to investigate gametocyte development and gene regulation thereof. To this end, a co-expression network was constructed to quantify the relationship between genes throughout gametocytogenesis and contrast this to processes required during asexual proliferation.

3.2 Methods

3.2.1 Parasite culture, sampling, and RNA isolation

Asexual *P. falciparum* NF54 parasite cultures (NF54-*pfs16*-GFP-Luc[156]) were maintained at 37°C in human erythrocytes at 5-8% parasitemia and 5% haematocrit in RPMI 1640 medium supplemented with 25 mM HEPES, 0.2 % D-glucose, 200 µM hypoxanthine, 0.2% sodium bicarbonate, 24 µg/ml gentamicin with 0.5 % AlbuMAX® II and incubated under hypoxic conditions (90% N_2, 5% O_2, and 5% CO_2)[115]. Synchronous asexual cultures (>95% synchronized ring-stage parasites) were obtained by three consecutive cycles of treatment with 5% v/v D-sorbitol, each 6-8 h apart. Gametocytogenesis was induced by through concurrent nutrient starvation and a decrease of haematocrit as described[95,115]. All cultures were maintained with daily medium changes and monitored with Giemsa-stained thin smear microscopy and parasite stage distribution determined by counting ≥100 parasites per day. Parasite samples (30 ml each of 2-3% gametocytaemia, 4-6% haematocrit) were harvested daily for RNA-seq analysis from two days prior to induction (day -2) and for 13 days post-induction. Parasites from samples harvested on days -2 to 7 were enriched via 0.01% w/v saponin treatment for 3 min at 22°C while samples from day 8 to 13 were enriched for late-stage gametocytes via density centrifugation using Nycoprep 1.077 cushions (Axis-Shield). All parasite samples were washed with phosphate-buffered saline before storage at -80˚C until RNA was isolated. Total RNA was isolated from each parasite pellet with a combination of TriZol treatment and using a Qiagen RNeasy kit (Qiagen, Germany) as before[94]. The quantity, purity and integrity of the RNA were evaluated by agarose gel electrophoresis and on a ND-2000 spectrophotometer (Thermo Scientific, USA).

3.2.2 Directional, amplification-free RNA-sequencing

3.2.2.1 DAFT-seq library

DAFT-seq was performed essentially as described in[25]. Poly-adenylated RNA (mRNA) was selected and captured using magnetic oligo-d(T) beads and purified. Full-length mRNA was reverse transcribed using Superscript II (Life Science), primed using oligo d(T) primers. Second strand cDNA synthesis used dUTP to encode directional information. The resulting cDNA was fragmented using a Covaris AFA sonicator. A "with-bead" protocol was used for dA-tailing, end repair and adapter ligation (reagents from NEB) using "PCR-free" barcoded sequencing adaptors (Bio Scientific). After 2 rounds of solid phase reversible immobilization (SPRI) clean-up the libraries were eluted in EB buffer and USER enzyme mix (NEB) was used to digest the second strand cDNA, generating directional libraries. Sequencing was performed on an Illumina HiSeq 2000 generating paired-end reads, at the Wellcome Trust Sanger Institute (WTSI), UK[25].

3.2.2.2 DAFT-seq primary data analysis

Reads were trimmed to 75 bp followed by alignment filter based read flagging. Reads that were PCR duplicates, mapped to multiple locations on the genome and improperly paired reads where discarded. Read mapping was done against the *P. falciparum* genome 3D7 version release 34 (www.genedb.org) using *TopHat2*[157], followed by minimum mapping quality threshold of 30 (*samtools* 0.1.19). Read count and FPKM normalization were calculated using in-house Bash and Perl scripts developed by Lia Chappell (WTSI). Read quantification and normalization was comparable to quantification of an alignment free method such as *Kallisto*[158]. Whole transcriptome Pearson correlations of sample time point where performed with filtering for strongly correlated values (Pearson $R^2 > 0.7$, with the exception of day -2: lower correlation due to low RNA abundance). The DAFT-seq dataset was evaluated for quality of reads using *fastqc* where per base sequence quality scores were primarily used to determine quality of reads. Post-read mapping to the *P. falciparum* genome the number of mapped reads compared to unmapped reads was used a coverage indicator.

3.2.3 Custom exploration of the gametocyte transcriptome

3.2.3.1 Whole transcriptome clustering and comparisons with established gametocyte datasets:

Previously published datasets were used to allow a full interrogation of both the intraerythrocytic development cycle (IDC) as well as gametocyte transcriptomes all from RNA-seq platforms to compare to the in-house generated DAFT-seq time course data. Datasets

were mapped from raw sequence reads as previously described. Genes with no signal/reads in more than 80% of the sample set were removed to ensure sufficient signal for normalization. Between sample normalizations of read counts were evaluated using a standard upper quantile normalizations (*DESeq2,*[159]), *Limma voom* [160] and variance stabilizing transformation (VST) (*DESeq2*).

The total transcriptome FPKM values from the DAFT-seq dataset as well as from the previously published datasets were scaled as z-scores and clustered both hierarchically and through K means clustering. Clusters were evaluated for their relevance during specific stages of the parasite. Various genes relevant to gametocyte maturation from known publications[21,95,105] were used to validate the dataset including a previous study of ours on gametocyte maturation (microarray). A panel of four genes were used in qPCR validation of the samples (*pf3d7_1252200, pf3d7_1302100, pf3d7_1104900* and *pf3d7_0525800*).

3.2.3.2 Weighted co-expression network analysis and intramodular hub gene identification

To identify sets of co-expressed genes during IDC and gametocytes stages respectively, the construction of a weighted gene co-expression network was performed using *WGCNA*[69]. Samples were evaluated for outliers by means of hierarchical clustering and soft threshold power coefficient was determined at 16 for scale-free topology (Supplementary Figure 1.1). From this adjacency matrix a topological overlap similarity matrix (TOM) was computed and the following dissimilarity matrix (1-TOM). Automatic block-wise module detection was used at a merge height of 0.3 and a minimum module size of 30 genes, which yielded 12 modules. In order to capture positively co-expressed genes, the network was created as a signed hybrid network. Singed hybrid networks consider only positively correlated genes. Uncorrelated modules (genes which showed no co-expression with other genes) were removed from that dataset for co-expression analysis. Manual curation of these genes showed their expression to be extremely variable over the datasets tested (Supplementary Figure 1.2). The remaining 11 modules were included in networks with layouts computed using the *igraph* package[161] Fructerman-Reingold algorithm for weighted network layouts of the strongest 5% edges. Visualization of networks were done using the *ggplots2*[162] in R and *igraph*. The strongest features (genes) for module identity (intramodular hub genes) were extracted for all 11 modules. Module eigengene correlations for each gene and the modular interconnectivity was calculated per module. The upper 5% for each were considered intramodular hub genes.

3.2.3.3 Module assignment to stage category data

Pearson correlations of module eigengene values (as defined in[69]) was used to categorise stage data into 4 broad stage categories: gametocyte (G), ring-trophozoite (RT), trophozoite-schizont (TS) and mixed gametocyte and asexual (Mix). A "one hot encoder" approach with binary encoding of the different categories ensured the categorical data remained non-continuous. Correlation of modules to stages was used to assign modules to their respective stages. A TS/G category was created for dual correlations where modules were equally correlated to two stage categories.

3.2.4 Stage assignment of genes with generalized linear modelling:

Genes were evaluated for stage-associated strength through generalized linear modelling (GLM). Genes were tested against each stage category and *P*-values for the total set of genes were adjusted (Bonferroni), p.adj<0.05. Predictor genes were evaluated for unique occurrence in stage (post-filtering) and visually inspected through heatmapping and hierarchical clustering.

Stage-associated genes derived from GLM were assessed for accuracy using a supervised machine learning approach. Four stage categories: ring-trophozoite (RT), trophozoite-schizont (TS), mixed (mix) and gametocyte (G) were used to predict the specificity of genes. Two sets of evaluations were conducted: testing using the full RNA-seq dataset used in the co-expression network (VST applied) and a second round of testing including two microarray datasets. For the second test, both RNA-seq and microarray data were transformed into -1, 0 or 1 depending on their presence in the upper, middle and lower quartiles of the datasets following mean centered normalization. This approach was used as decision tree algorithms are sensitive to positional rankings of the data and RNA-seq with microarray are not directly compatible with regards to exact transcript abundance positions. The model was trained using scikit-learn *GradientBoostingClassifier* algorithm with 5000 estimates. Data was split into a 70:30 training-test ratio and 10x cross validation was used to assess model performance on the training set using *cross_val_predict* from scikit-learn. Test sets were evaluated simply for accuracy.

3.2.5 Text mining and filtering of regulatory machinery from current annotations

Nuclear genes were cross-referenced for gene ontology terms and assigned to each gene, terms sourced form PlasmoDB (www.plasmodb.org) and ApicoTFdb[146] for added information on transcription factor elements. Unknown proteins from this list was also submitted to InterProScan[163] via a programmatic interface with the RESTful API service to further cross-reference any potential regulatory elements overlooked by current annotations. Text filtering using regular expression on customized search terms were used to filter genes into three categories: translation machinery, transcription machinery and epigenetic machinery. Genes were subsequently cross-referenced with modules in the co-expression network in order to polarize the machinery mechanisms throughout the dataset.

3.2.6 Evaluation of intergenic long non-coding RNA (lncRNA) and adjacent neighbouring gene pairs through co-expression data

For evaluation of lncRNA we required paired-end datasets and a modification to the correlations statistic to capture any negatively correlated relationships as lncRNA may exhibit repressive or expression interfering properties. Two datasets were kept for this analysis as their library preparations were compatible for the analysis, 4 datasets were removed due to either unpaired reads or unsuitable library preparation. The modification to the correlation statistic is through means of *csuWGCNA*[164] network which takes absolute values of correlations between pair thus maintaining negatively correlated pairs (equations 1-2).

[1] $Signed\ aij = |(1+\text{cor}(xi,xj))/2|^{\beta}$

[2] $csu\ aij = |(1+|\text{cor}(xi,xj))|/2|^{\beta}$

Where *a* denotes adjacency values (weighted correlations) with *i* and *j* for respective gene pairs. Cor denotes correlations and $^{\beta}$ denotes the power coefficient required to achieve scale-free topology. Furthermore, the predictive nature of lncRNA for correlated nuclear genes were evaluated through use of generalized linear models (base R: *glm*), and Bonferroni adjusted P-values were calculated. The lncRNA data was filtered for weighted correlations > 0.4 and adjusted P-values < 0.001 producing a subnetwork of lncRNA genes and their subsequent target genes. Neighbouring gene pairs were filtered into three configurations: Tail-to-head, Head-to-head and Tail-to-tail for matched gene pairs with *csuWGCNA* values assigned. For the identification of intragenic antisense RNA (asRNA), our RNA-seq dataset was probed for

asRNA and a log_2FC (asRNA + 1/mRNA + 1) was calculated and clustered (k-means, with k = 9).

3.3 Results

3.3.1 High resolution transcriptome from DAFT-seq exhibit clear phase differentiation in the parasite development:

3.3.1.1 RNA-seq read quality and mapping

RNA sampling was successfully performed to allow subsequent deep sequencing of the transcriptome. This was performed for all 16 samples spanning asexual parasites and commitment to gametocytogenesis and days 1-13 of gametocytogenesis for all stages of gametocytes. A base pair read length of 75 bp was submitted to FastQC for each of the 16 samples. The base pair quality per mapped reads is classified as high-quality for all the samples evaluated (average quality scores of >28) (Figure 3.1A). Base pair composition for each read is shown as an average for all mapped reads over all samples in Figure 3.1B. As expected, the GC content of the reads are relatively low, mimicking the parasite genome's low GC content (~19%)[165]. Each sample had over ~80% mapped reads with most samples at ~90% (Figure 3.1C).

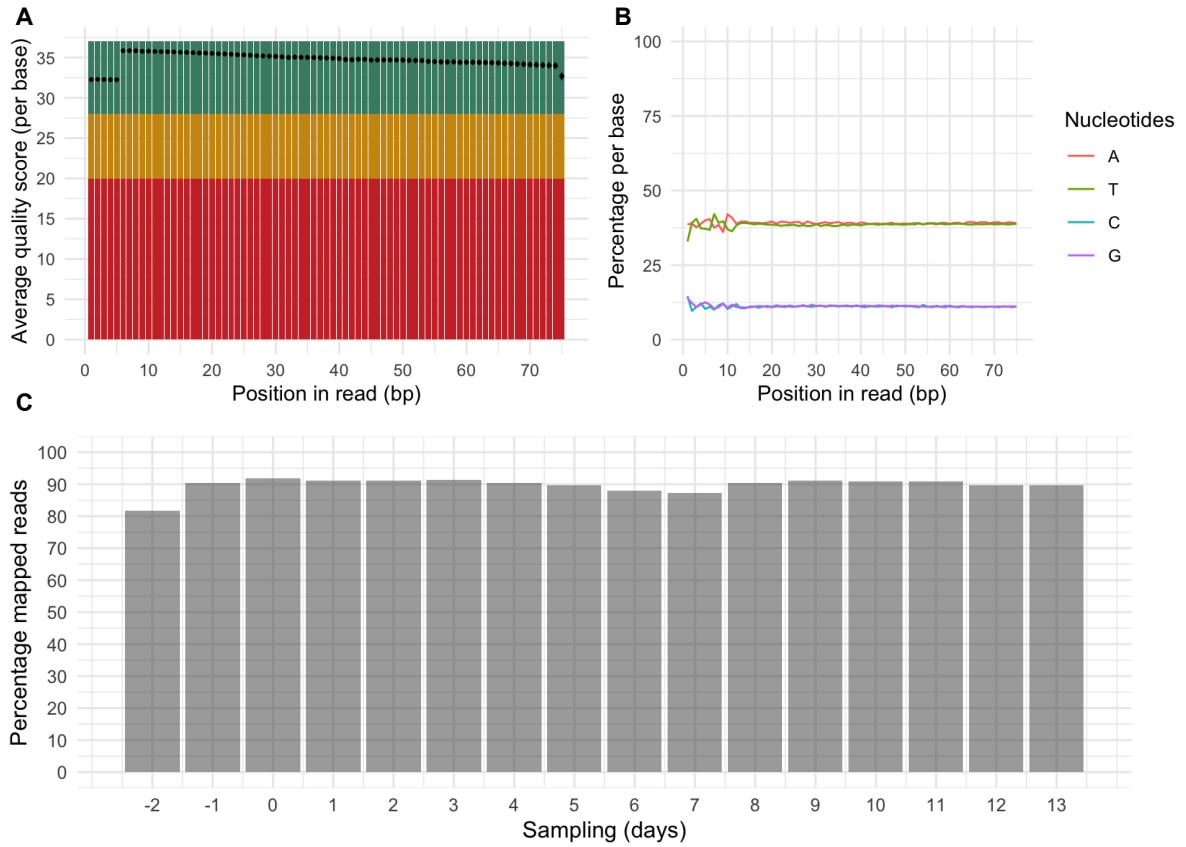


Figure 3. 1: RNA-seq quality statistics.
A) Average base pair quality scores for all mapped reads across all samples as produced by FastQC reports. Quality ranges are colour coded: green = high quality with average quality scores >28, yellow = moderate quality with scores >20 and red = low quality <20. Where standard deviation not visible, it falls with the datapoint symbol. B) Average base pair composition for all samples, all four nucleotides are indicated. C) Percentage mapped reads per sample.

3.3.1.2 Sample stage composition and sequence distribution

Gametocytogenesis, development and differentiation were morphologically monitored over a 16-day period that covers induction of asexual parasites for commitment to gametocytogenesis (day -3), to the first appearance of stage I gametocytes (day 1) through to mature stage V gametocytes on day 13 (Figure 3.2A). Samples for DAFT-seq were taken daily during this entire time course and stage composition determined to recapitulate expected profiles in each stage (Figure 3.2A)[95]. Morphological evaluation of parasite stage composition at day 3 show an 81% gametocyte population and 97% at day 4, whereafter only gametocytes were present in various stages from day 5 onwards (Figure 3.2B). Moreover, from days 3 to 13, the gametocyte stage-distribution was evident with maturation from stage I gametocytes on day 3 progressing to a majority stage II population by day 5, and stage III 24 h later. Late-stage gametocytes were evident from day 7 and by day 13, gametocytes were mostly mature stage Vs. The dynamics of this progression of gametocyte development was reproducible over

replicates and mimicked in its entirety a previous profile[95] on this strain of *P. falciparum*. This provided confidence in the stage coverage provided with our sampling strategy.

The subsequent global transcriptomes obtained for each sample from the DAFT-seq strategy shows a comparable sequence coverage distribution throughout the 16 samples suggesting equal extraction and sequencing of transcripts during isolation (Figure 3.2C). The distribution of mapped reads throughout the dataset exhibits a clear comparative range between samples with only trophozoites illustrating a higher range in distribution as expected. Pearson correlation between the DAFT-seq data from the daily sample sets confirm the morphological distribution, with days 3-4 showing the highest heterogeneity, whilst pure gametocyte populations corresponding to intermediate and late-stage development is clearly separated from asexual development, with a strong sample clique between days 5-13 (Figure 3.2D).

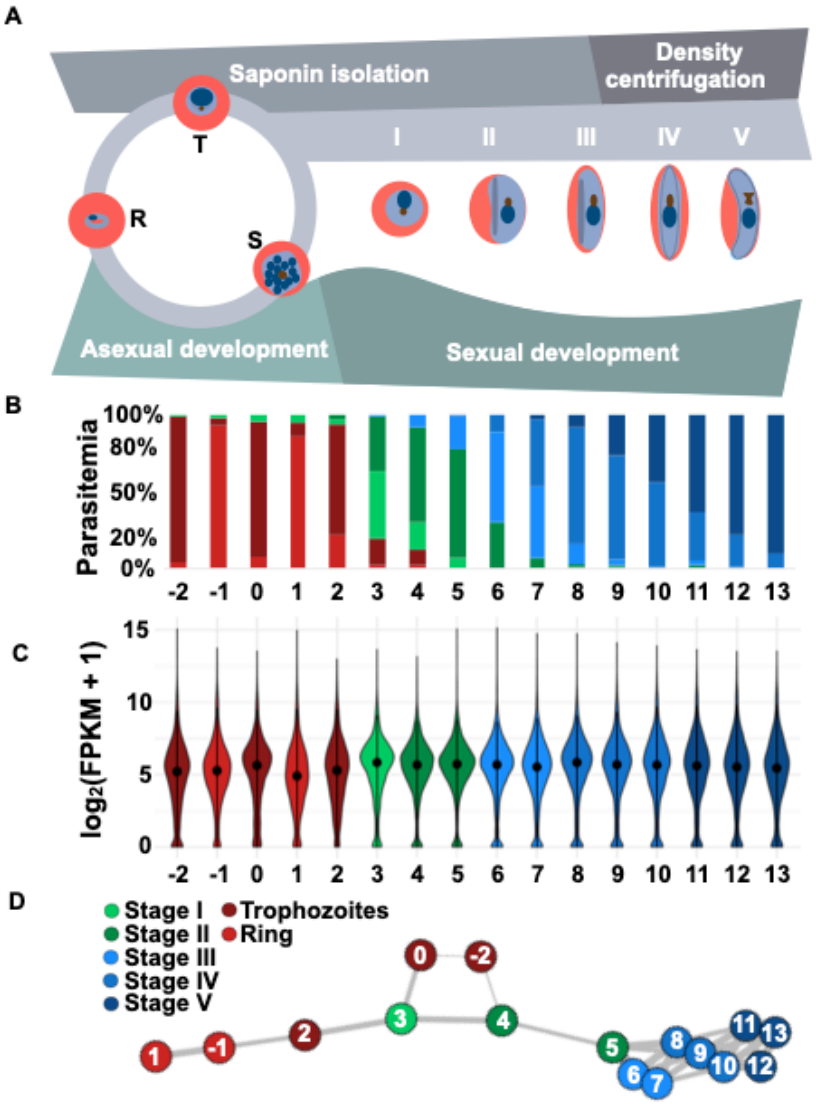


Figure 3.2: Inclusion of high-resolution gametocyte development transcriptome through RNA-seq based platforms.
A) Sampling and culturing strategy for in-house DAFT-seq transcriptome capturing induction of gametocytes to maturation. Daily sampling done with induction started at day -3 and sampling from day -2 to 13 and parasites binned into stages based on Giemsa stained microscopic evaluation[95]. B) Stage composition of daily sampling from in-house dataset from day -2 to 13. C) Distribution of log_2(FPKM + 1) values for each daily sampling with indicated median. D) Whole transcriptome sample Pearson correlations represented as a network with nodes indicating sample day and edges the correlation strengths. Figure colour legend indicates parasite stage: R = ring stages, T = trophozoite, S = schizont and gametocyte maturation indicated as roman numerals I-V.

3.3.1.3 Gametocyte signatures present strongly in the sampled transcriptome

The DAFT-seq transcriptomes for the full gametocyte developmental programme was compared to previous cDNA microarray transcriptomes[95]. Similar deviation in the entire transcriptome from asexual parasites to gametocytes were observed in the DAFT-seq transcriptome compared to our transcriptomic data from cDNA microarrays[95]. We identified 583 transcripts that showed uniquely and specifically increased expression profiles during gametocyte development in comparison to asexual parasites (Figure 3.3).

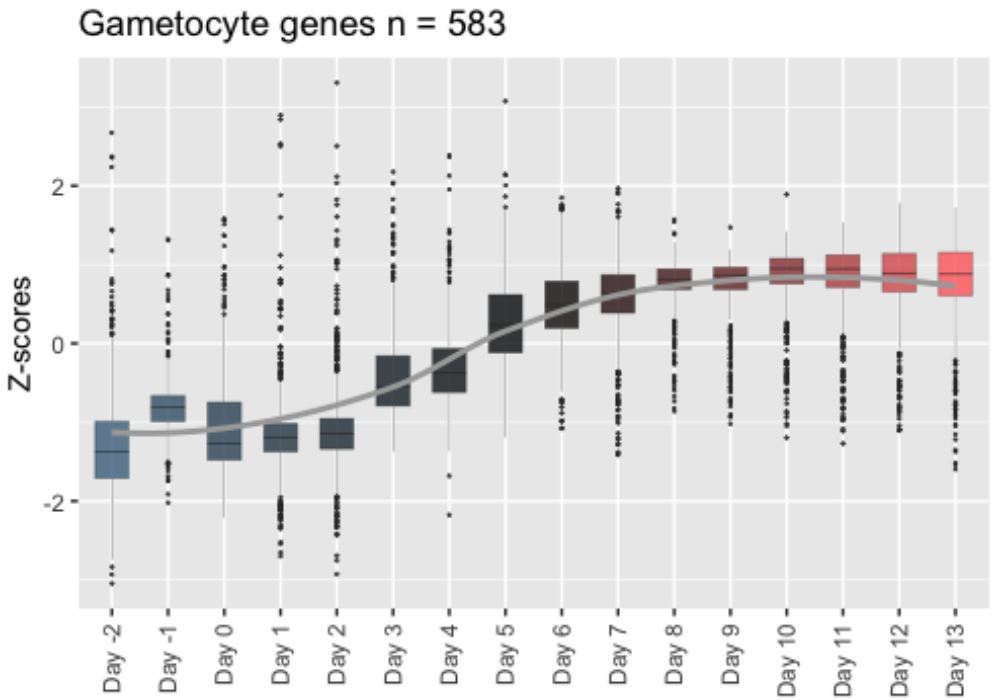


Figure 3. 3: Gametocyte associated genes reveals expected expression trends throughout the time course data.
A sample of 583 genes were established to be gametocyte associated genes[95] and the expression mean was modelled using loess fit. Sample distribution of genes are illustrated as box-whisker plots. Box colour ranges shows progression of sampling through time points, from blues (asexual samples) to reds (mature gametocytes). Expression values are treated as Z-scores of FPKM.

In addition, our confidence in our ability to pick out gametocyte-related transcripts is also bolstered by identifying 37/40 “gold standard” gametocyte genes[21] that show increasing in transcript abundance in our RNA-seq dataset (Figure 3.4). Three of the 40 genes represented in Figure 3.4 (*pf3d7_1302100, pf3d7_1477300* and *pf3d7_1253000*) are present during early stages of gametocyte development only. The concordance with previously established gametocyte findings here recapitulates the fidelity of the DAFT-seq transcriptome dataset as a strong representation of gametocyte development in the parasite.

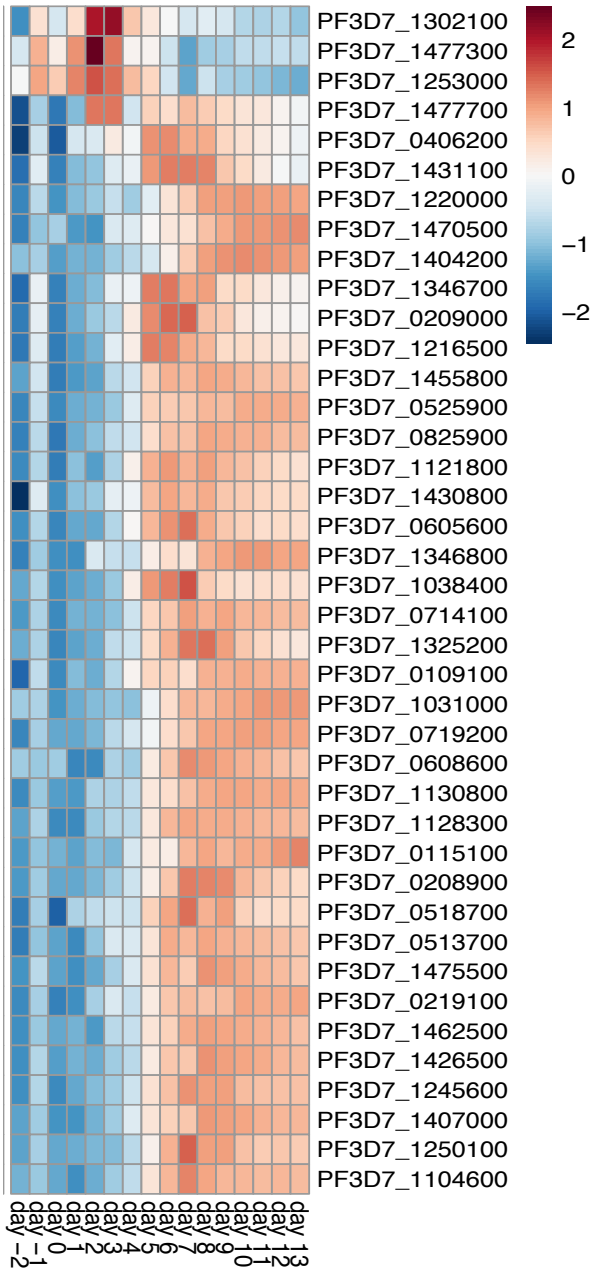


Figure 3.4: Meerstein-Kessel gold standard of sexual development genes for the in-house RNA-seq dataset.
A total of 40 genes represented.

These data show consistent gametocyte sampling and the presence of strong gametocyte signals in concordance with previous studies, supporting subsequent mining of this dataset. Typical advantages of RNA-seq over that of microarray data, is the potential discovery of novel ncRNA and the investigation of alternative splicing, although the latter is not typically easily investigated with base pair reads of 75 bp. TSS discovery, however, requires a more stringent sequencing process that protects the 5' ends of the transcript, which the DAFT-seq applied here does not perform and therefore TSS discovery could not be applied.

3.3.2 Normalisation approaches across numerous datasets yield variance stabilising transformation (VST) as the most appropriate

Given the availability of RNA-seq dataset for *P. falciparum*, integration of data would lend greater statistical power in downstream analysis between the developmental phases. Samples in addition to our gametocyte time course dataset were included from public repositories such as PlasmoDB.org, EMBL and GEO. Data from these repositories were chosen based on several criteria: 1) the data had to be from bulk RNA-seq experiments; 2) it had to include different parasite stages in both sexual and asexual phases of development. Numerous IDC RNA-seq datasets were available for use, however not many bulk RNA-seq datasets exist for gametocytes and the available gametocyte datasets are for only specific stages, never for the entire spectrum of gametocyte development. There was therefore a heavy bias towards inclusion of RNA-seq data from asexual stages. Ultimately, data from 49 samples were integrated into a complete dataset (Table 3.1) to allow the construction of a weighted co-expression network.

Table 3. 1: Datasets used for co-expression network with a total of 49 samples through all datasets.

Dataset description	Abbreviation	Library type	Sample#	Reference	Access
Hoeijmakers IDC (5-40 h)	HM	Unpaired-end	8	-	(PlasmoDb.org)
Siegel IDC (10-40 hr)	SL	Paired-end	4	[153]	ERP001849
Broadbent IDC (8-40 h)	BB	Paired-end	14	[43]	GSE57439
López-Barragán (S,LT,GII,GV)	LB	Paired-end	4	[149]	(PlasmoDb.org)
Lasonder (GV)	LS	Unpaired-end	3	[22]	GSE75795
Van Wyk (R,T,GI,GII,GIII,GIV,GV)	RW	Paired-end	16	In-house	-
Total			**49**		

S: schizont, LT: late trophozoites, R: ring stages, for GI-GV: G = gametocyte and I-V indicate stage.

Cross-sample comparisons are only feasible following appropriate normalisation of the dataset to create comparable data ranges between samples in the entire dataset. Four main normalisation strategies were evaluated to successfully integrate the data from the various sources (Figure 3.5). A $\log_2$ transformation of read counts are included as a normalisation reference point, the real comparison focuses on upper quantile (UQ), *limma's* voom and VST

from *DESeq2*. Upper quantile normalisation, which is considered as a library size normalisation strategy, produced a widely varying median across samples sets observed. Transformations with either *limma* voom or VST are usually recommended for co-expression network analysis[69], and both shows comparable normalisation but voom does have less uniformity between samples. For this reason, VST appears to be the more appropriate tool for read count normalisation between samples.

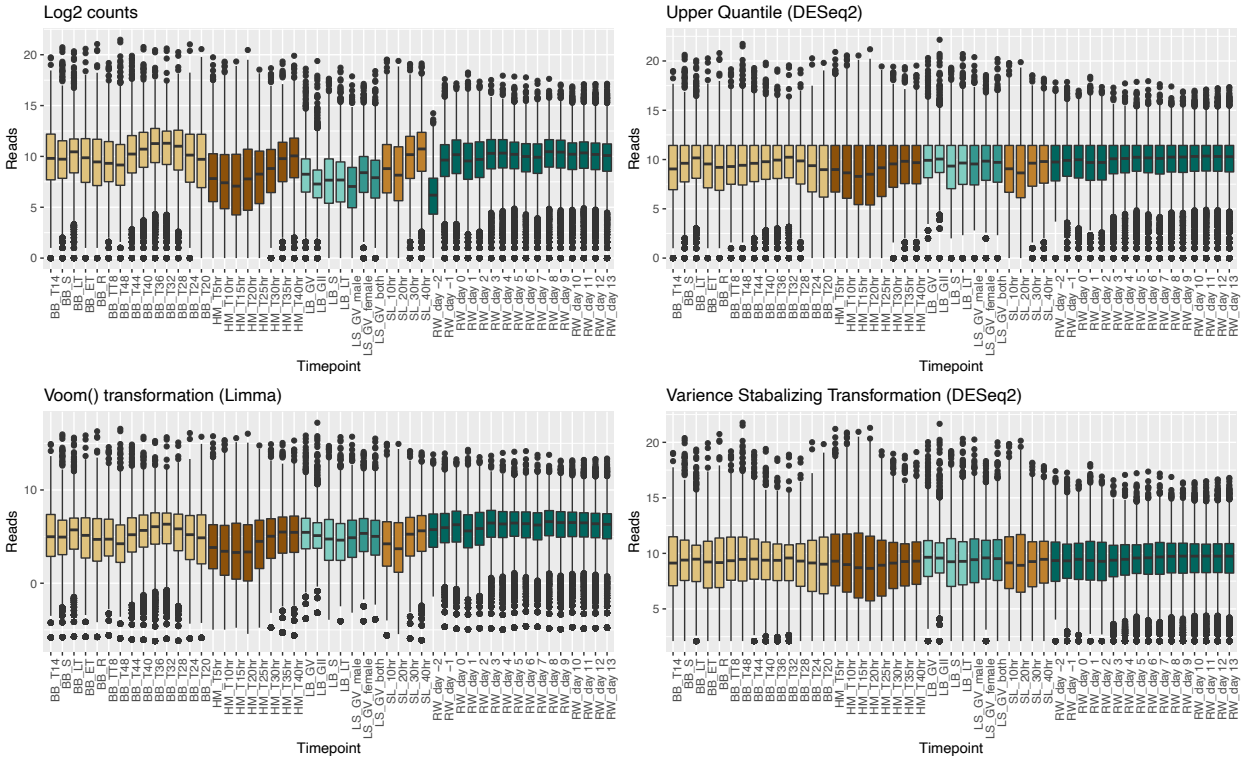


Figure 3.5: Dataset normalisation strategies.
Log_2 transformation, upper quantile normalisation, limma voom and VST from DESeq2. Sample keys: BB = Broadbent, HM = Hoeijmakers, LB = López-Barragán, LS = Lasonder, SL = Siegel, RW = van Wyk. Each dataset is illustrated by a unique and corresponding colour: beige = BB, dark brown = HM, turquoise = LB, sea green = LS, orange = SL and dark green = RW

3.3.3 Correlation of the total dataset samples yield clear stage categories for general use and assignment of clusters

Substantial variance is observed in data collected for '-omics' experiments from *P. falciparum* due to various experimental conditions used in each dataset (e.g. different sampling times and sequencing platforms). Variance of this nature make downstream interpretations difficult, whilst correcting for this variance through normalisation may improve interpretations. We therefore clustered the normalised data based on the stage of parasite development and total transcriptome correlation. We defined broad, stage-associated categories as follows: ring-

trophozoite (RT), trophozoite-schizont (TS), gametocytes (G) and mixed (Mix) populations. This resulted in four clear correlation blocks (black boxes) associated with the four categories (Figure 3.6).

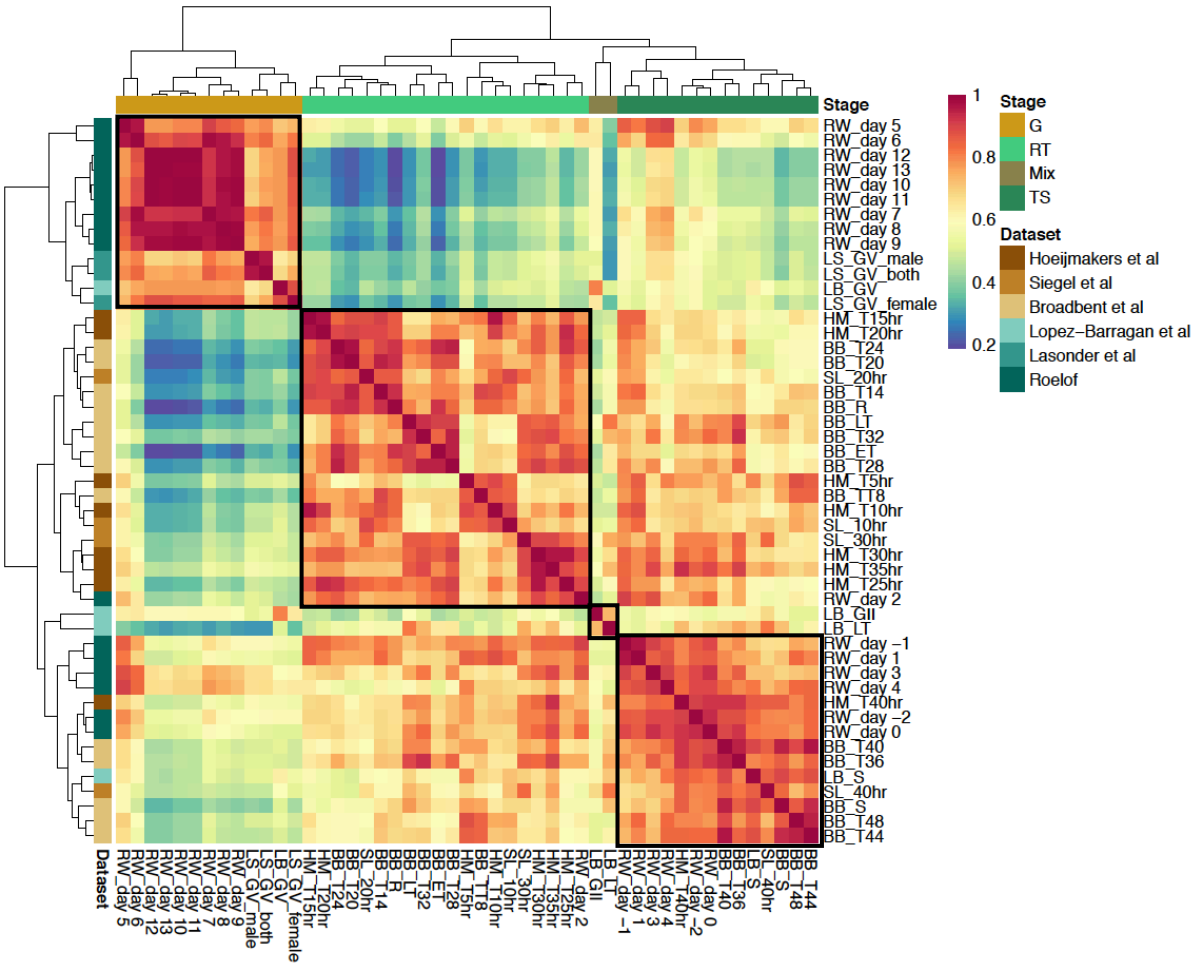


Figure 3. 6: Sample severances through Pearson correlations classified within developmental blocks.
Definition of stage categories based on sample correlations blocks (black boxes) and prior knowledge of sample composition. Stage category labels: G = gametocyte, RT = ring-trophozoite, TS = trophozoite-schizont and Mix = mixture of gametocyte and asexual parasites. Datasets reference in legend. Sample normalized using the VST (DESeq2) approach in R. Sample keys: BB = Broadbent, HM = Hoeijmakers, LB = López-Barragán, LS = Lasonder, SL = Siegel, RVW = van Wyk

Interestingly, the gametocyte stage II and late trophozoite stage from the López-Barragán dataset formed a correlation block which was more strongly associated with the TS stage samples of the other datasets. This is most likely due to a mix of asexual and sexual stages present in these samples. Mixed sample populations (particularly LB_LT, LB_GII and days 3-4) introduced a significant amount of noise in the dataset but this is considered fortuitous in terms of stage-category assignment of genes. Noise reduction by removal of mixed sample populations will increase the likelihood of a type I error, while inclusion skews towards a type II error. Since we are interested in discovering genes which most significantly assign to a

particular stage-associated category, type II errors are generally more acceptable. Samples within the TS block may very well contain early-stage gametocyte samples, however, the signal from these gametocytes is expected to be heavily masked by the number of asexual parasites populating each time point. A clear gametocyte block is seen for late-stage gametocytes (GIII to GV) that is completely dissociated from all of the other sample categories. A large RT stage block is present in the correlation matrix. Formation of clear correlation blocks removes some of the biases in assigning samples to specific stages across multiple datasets. Downstream interpretations must be mindfully performed with these stage categories.

After normalisation supported clear stage separation in sampling across multiple datasets, further probing of the data can be performed. The proposed method for producing more insights into the relationships between genes, was through the construction of a WGCNA GRN.

3.3.4 WGCNA highlights a strong bimodal distribution of co-expressed genes for sexual and asexual development in *P. falciparum*

Validation of gametocyte sampling both morphologically and transcriptomically, secures a large contributing dataset towards gametocyte gene expression with RNA-seq. Since our dataset includes all morphologically distinct stages of gametocyte development, we can include this with existing data to improve the statistical power of the downstream analysis. Integration of these data will enable clear developmental stage categories spanning both asexual and sexual development to be delineated but requires a global approach which can segment the data into relevant groupings for multiple comparisons. We therefore constructed an undirected, weighted co-expression network with WGCNA as a first attempt at a global GRN. WGCNA also includes clustering approaches, which segment the data based on expression signatures and captures the relationship between transcripts through weighted Pearson correlations.

The total transcriptome for all 49 samples (5163 transcripts shared for each dataset) were used to construct the WGCNA, during which module discovery through clustering of correlated genes separate the transcripts into distinct modules. A total of 11 modules were discovered that spanned 4100 transcripts and covering 75% of the transcriptome. Transcripts that were excluded from the network showed no association with any of the modules and were denoted unsigned. The top 5% with strongest weighted Pearson correlations (n=3306 transcripts) were

used to capture the network topology with the Fructerman-Reingold algorithm (Figure 3.7A). The network presents with a bimodal distribution with some localised and some widely dispersed modules (Figure 3.7A). The WCGNA indicated the presence of intramodular hub genes that show high connectivity throughout the network and present with a centralized topology within their respective nodes. These nodes can therefore be considered as the most representative subsamples of their respective modules.

Interestingly, the connectivity associated with the intramodular hubs resulted that these hubs, with their strongest interacting nodes, form a network "backbone" of 774 genes (Figure 3.7B). Extremely dense cliques form throughout this "backbone" network with modules 2 & 3 and 5-7 more densely compacted and clearly separated from modules 1 & 4. Like the clustering data, most of the genes associated with modules 1 and 4 as is reflected in the size of these modules. However, this area of the bimodal network that contains modules 1 & 4 also overlaps with modules 9-11, which contained very few genes. The backbone network preserves the general topology of the overall network, which speaks to the scale-free topology nature of the network itself and confirms that this network for *P. falciparum* remains relatively constant as is often the case with biological networks [73], and is independent of gene set sizes.

Stage category association of the modules was subsequently done using Pearson correlations of the modules with the respective staged categories (mixed, RT, TS or G; Figure 3.7C). Taking module eigengene correlations to stage categories into account, dual nature co-expression was observed for some modules, with similar correlations occurring in more than one stage category (Figure 3.7C). However, distinct associations could be inferred with modules 1, 4 and 8-10 showing strong eigengene correlations (≥ 0.6-0.7) to stage categories associated to asexual RT stages, with some overlap present particularly for module 9 for TS. The distinction of these modules from the others implies a clear asexual pole defined by the WGCNA, confirming previous indications of deviation of the asexual and sexual transcriptomes[94,95]. This is extended to module 5 that is associated with TS and could provide the strongest representation of schizont data in this dataset. Importantly, this module is also associated with a bridging nature between the poles that could indicate transition from asexual parasites to gametocytes.

Although a less pronounced correlation was observed with modules associated to gametocyte stages (≥0.3-0.4, module 7, 6 and 2), module 3 showed very strong and exclusive association with gametocytes (≥0.93, Figure 3.7D), confirming that a subset of genes differentiates and describe gametocytes[95]. These genes also show very high connectivity, implying functional relationships or co-regulation. To further distinguish gametocyte-associated genes and define

possible overlap with asexual-related genes, a separate category was defined termed TS/G since the correlation of modules 2 and 6 were relatively equal for both TS and G stage categories (Figure 3.7D). This finer delineation showed that for these genes, multi-stage involvement can be assigned, with little distinction between them in mature asexual stage parasites (TS) compared to gametocytes. This may indicate either shared importance to both these stages or could reflect overlap in populations of sexually committed asexual parasites and immature gametocytes.

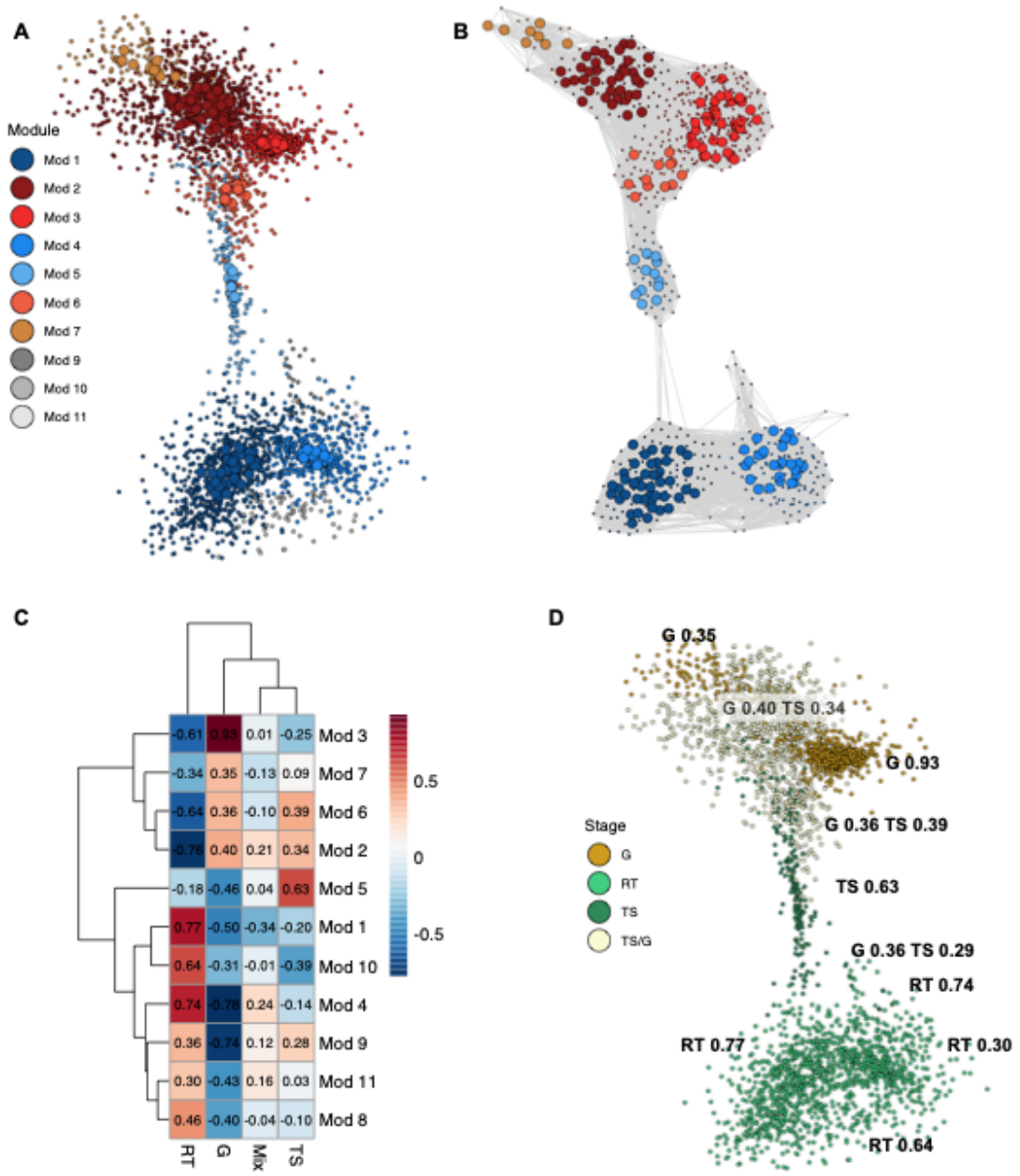


Figure 3. 7: Weighted co-expression network analysis captures a bimodal distribution of co-expression in developmental stages.
A) Co-expression network with imposed gene modules (colour scale) illustrating a bimodal distribution of co-expressed genes. Visualization of the top 5% strongest interactions (n = 3306 out of 4158 genes). Larger nodes correspond to hub nodes. B) Hub genes throughout modules and their strongest interacting genes forming a network "backbone" of 774 genes. Hub nodes are of larger size than non-hub nodes, colour scale indicate module membership. Edges show interactions in network (not scaled). Nodes were positionally declutter using a coordinate offset for clearer visualization. C) Correlation matrix of module eigengene values and respective stage categories. D) Correlations of modules to sample stage category imposed in the network topology. Stage category legend: RT = ring-trophozoite, TS = trophozoite-schizont, G = gametocyte and TS/G = ~equal correlations to either TS or G categories. Correlations for modules indicated within proximity of their topology.

The functional outcomes of these modules were subsequently evaluated using gene ontology (GO) enrichment. Pie scatter plots (placed at module centroids) indicate enriched GO terms

found throughout the network topology (Figure 3.8). At the core of the gametocyte assigned gene modules, module 3 principally associated with the expected GO terms such as microtubule-based processes/movement (GO; GO:0006928; P=3.602e-10), lipid biosynthesis (GO; GO:0006633; P=1.514e-4) and cell motility (GO; GO:0048870; P=9.198e-4). These processes are all important to mature gametocytes in preparation for gamete formation and egress[95,105]. Modules 5 and 6 are both enriched for host cell entry processes (GO; GO:0030260; P=4.631e-20 and P=1.913e-3), that includes expression of transporters and surface antigens required for asexual proliferation processes including immune evasion and erythrocyte entry[94]; but also needed for host cell remodelling and entry for immature gametocytes to sequester in tissues such as bone marrow to allow differentiation to occur[166].

By contrast, module 1 which associates with RT stage categories show enrichment for asexual-related processes including RNA metabolic process (GO; GO:0016070; P=1.640e-11), ribosome biogenesis (GO; GO:0042254; P=5.460e-10) and RNA processing (GO; GO:0006396; P=2.820e-10)[49]. The trophozoite-selective nature of module 2 is heavily reflected in the GO terms, with these pertaining almost exclusively to DNA replicative processes (GO; GO:0006260; P=1.369e-23). The ability of the network to segment the data into relevant modules within respective stage categories, as further supported by the underlying biology associated with these stages, provides confidence in the nature and accuracy of the network. It could therefore be used to delineate stage-specificity, provide chorological order to gene expression, and evaluate the regulatory importance of the connectivity.

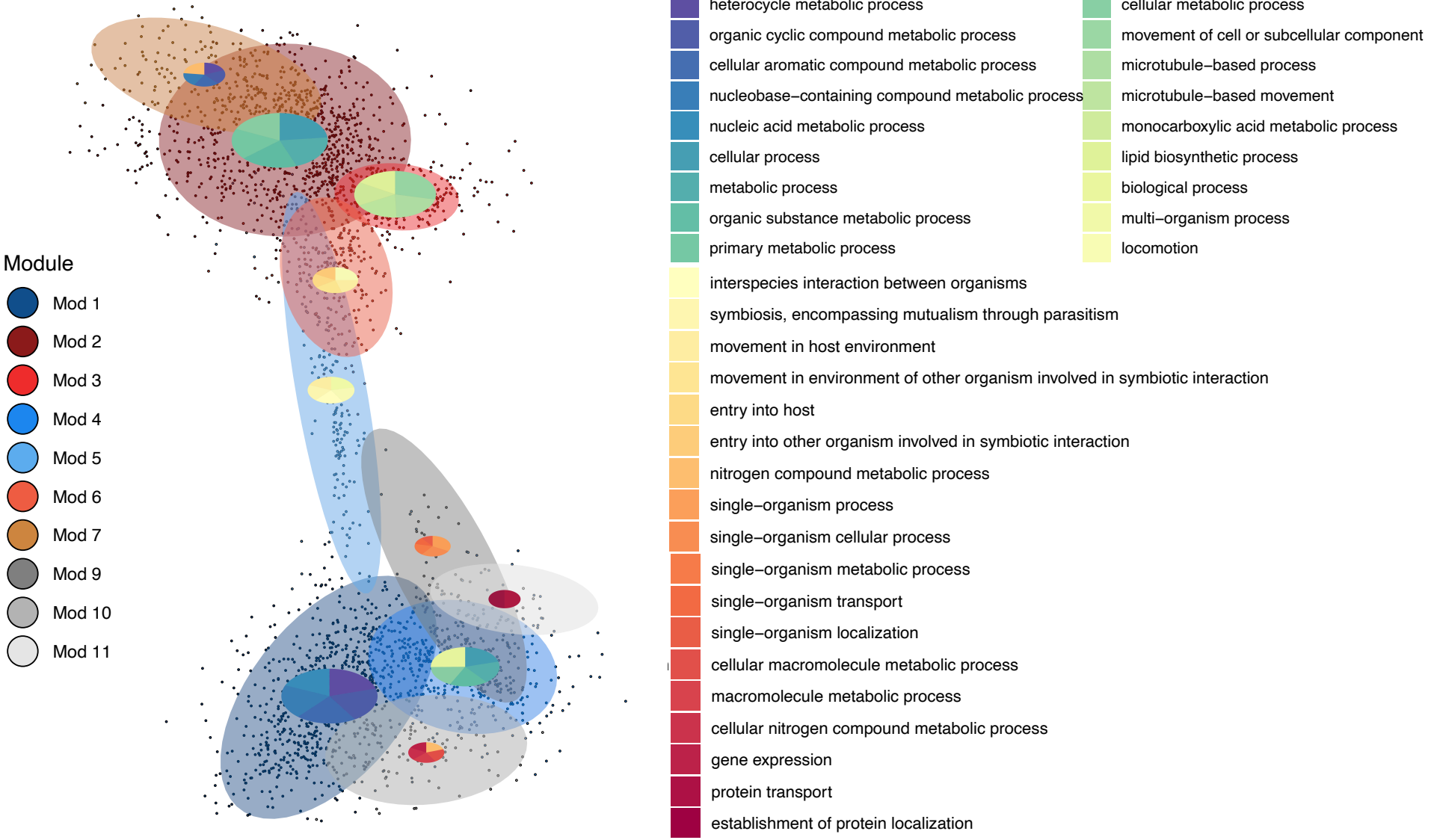

Module
Mod 1
Mod 2
Mod 3
Mod 4
Mod 5
Mod 6
Mod 7
Mod 9
Mod 10
Mod 11
heterocycle metabolic process
organic cyclic compound metabolic process
cellular aromatic compound metabolic process
nucleobase−containing compound metabolic process
nucleic acid metabolic process
cellular process
metabolic process
organic substance metabolic process
primary metabolic process
cellular metabolic process
movement of cell or subcellular component
microtubule−based process
microtubule−based movement
monocarboxylic acid metabolic process
lipid biosynthetic process
biological process
multi−organism process
locomotion
interspecies interaction between organisms
symbiosis, encompassing mutualism through parasitism
movement in host environment
movement in environment of other organism involved in symbiotic interaction
entry into host
entry into other organism involved in symbiotic interaction
nitrogen compound metabolic process
single−organism process
single−organism cellular process
single−organism metabolic process
single−organism transport
single−organism localization
cellular macromolecule metabolic process
macromolecule metabolic process
cellular nitrogen compound metabolic process
gene expression
protein transport
establishment of protein localization

Figure 3.8: Gene ontology (GO) for network modules capture through pie scatter plots and module density shown with ellipses.
Pie scatter positionally placed at module topology centroids, therefore the centre where modules are most dense. GO terms are colour scaled as per legend description. Ellipses graphically illustrate module density in two-dimensional space. Only GO terms with result count>10 and *P-value*<0.05 were considered (supplementary table 1).

Given stage associated module and GO analysis validations, we can further examine the fidelity of the network through comparisons as done for the RNA-seq dataset previously[21,48,95,105]. A comparison of the Meerstein-Kessel "gold standard" of genes associated with specific developmental phases, provides this added validation of module fit[21]. All of the previously described 'gold standard' gametocyte-related genes are associated with module 3 (Figure 3.9A), with only one gene, an annotated gametocyte-exported protein (*pf3d7_1253000*) not expressed in mature gametocyte stages but rather captured during early asexual development (RT stage category) [167]. This was also observed in other gametocyte transcriptomes such as van Biljon *et al.* and Young *et al.,*[95,105].

The asexual 'gold standard' markers were associated with expression in asexual modules 5, 4, 1 and 8. Two genes (*pf3d7_1216600* and *pf3d7_1418100*), showed higher expression during gametocyte stages (Figure 3.9B) as well as association to gametocyte modules, implying that gene stage-association may therefore not be entirely solved. Barring these exceptions, these data support the ability and topology of the GRN to correctly assign stage association of genes. The question of gene stage association was however prompted by these exceptions and whether we need to refine our understanding of stage association for these data. Given that we would have to interpret the data in context of which stage of development this gene seems to more or less associate, evidence of the individual gene associations (not just group through model assignment) would be valuable for interpretations.

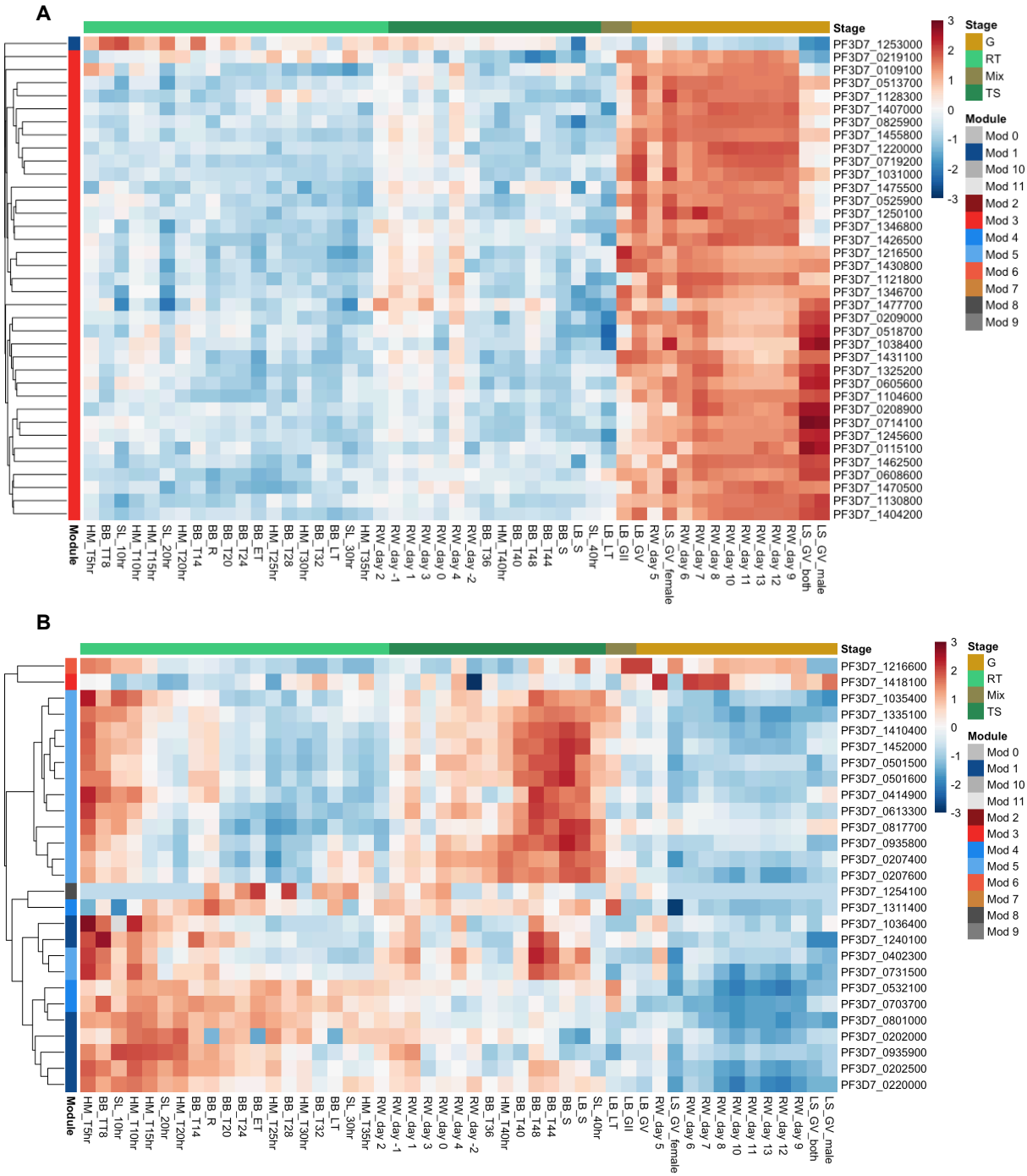


Figure 3.9: Gold standard asexual and sexual gene sets from Meerstein-Kessel compared to co-expression network:
A) Sexual stage gold standard set compared. B) Asexual stage gold standard set compared. Expression repressed as z-score values.

3.3.5 Generalized linear modelling (GLM) reveals predictors of stage predominant genes

Given the noted discrepancies between the known 'gold standard' for developmental phases, an analysis was extended to the entire dataset to evaluate transcript stage association. GLMs were created to identify the stage for which the transcripts are most predictive based on linear regression. Previously defined stages (RT, TS and G) were used to group transcripts within each category and GLMs were applied before filtering the results for Bonferroni adjusted *P*-value < 0.05 (Figure 3.10). A total of 1420 genes where significantly assigned to one of the defined stage categories and this included 500 (~35%) previously uncharacterized genes in the set. Genes found to be significantly associated with the RT stage category totalled 614 (~43%) genes, 119 (~19%) of which are uncharacterized. Four ApiAP2 transcription factors were significantly associated with the RT category: *pf3d7_1139300, pf3d7_1466400* (*ap2-exp*), *pf3d7_1408200* (*ap2-g2*) and *pf3d7_0730300* (*ap2-l). Ap2-g2* was recently shown to associate with RT stages as a general repressor of gametocyte-related genes during asexual proliferation[41], supporting the accuracy of our predictions. The stage specificity, however, does not show exclusion of the transcripts during other stages in development. For example, *ap2-g2* also express during early-stage gametocytes, but not as much as during asexual stages[95]. Of the intramodular hub transcripts, 73 (~60% of stage hubs) were found to be RT associated. Serine/threonine protein kinase (*pf3d7_1230900*) and bromodomain protein (*pf3d7_1475600*) are amongst these transcripts (Supplementary 2).

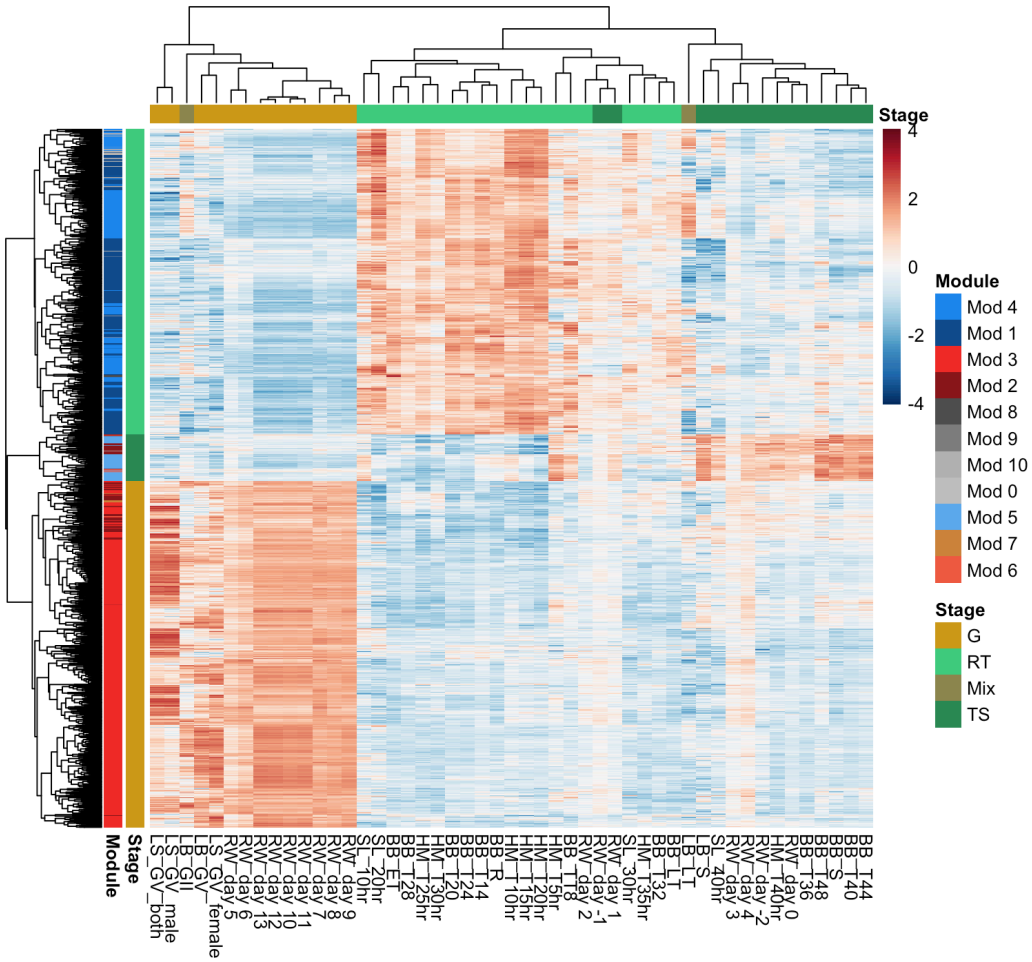


Figure 3.10: Generalised linear model output of stage-associated genes.
A total of 1420 genes where significantly associated to stage categories. Expression represented as z-score values. Data available in chapter 3 supplementary 2.

For the TS stage category, a total of 93 genes (~6.5%) was significantly associated with this stage, with 27 (~29%) uncharacterized genes. Four ApiAP2's was associated with this stage: *pf3d7_1239200, pf3d7_1456000, pf3d7_0613800, pf3d7_0604100*. Only 10 (~8.3%) transcripts form the intramodular hub set associated with this category. Stage associations to gametocytes accounted for most of the gene associated sets at 713 (~50%) genes. The uncharacterized genes in the gametocyte set accounted for ~50% of the gametocyte genes. Only one ApiAP2 was significantly associated to the gametocyte stages: *pf3d7_1429200* (*ap2-o3*), recently shown to regulate sex-specific expression of female genes in gametocytes[168]. Intramodular hub genes (38 or ~31%) also associated with the gametocyte stages. Strong markers of late-stage gametocytes are captured in the intramodular hub subset, with genes such as *psop13* (*pf3d7_0518800*), *apl3* (*pf3d7_0728200*), ankyrin-repeat protein (*pf3d7_0825100*), WD-repeat proteins (*pf3d7_1104500* and *pf3d7_1121400*) and ULG8 (*pf3d7_1234700*). Stage-associated genes were cross-referenced with the known "gold standards" as set by[21] and seen in Figure 3.11, which shows strong concordance with the

known published data. The GLM results are based solely on transcriptomic data, this makes the data highly suitable for transcript quantification planforms such as qPCR. These stage-associated transcripts will be referred to as the stage-panel transcripts.

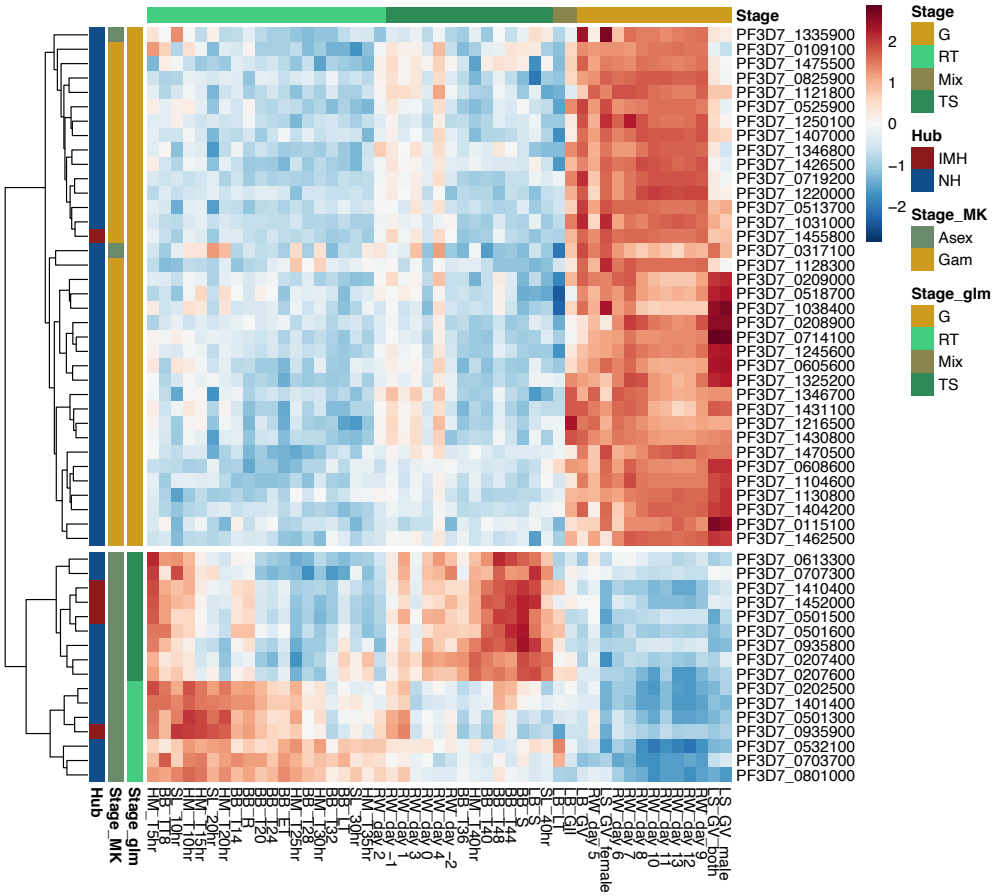


Figure 3.11: Meerstein-Kessel gold standard cross-referenced with GLM staged genes and intramodular hub genes.
Stage_MK = Meerstein-Kessel stage description. Stage_glm = our stage description. Hub IMH = Intramodular hub genes, NH = non-hub genes. n = 52

The stage-panel transcripts sets were further evaluated using supervised learning approaches to determine the predictive power of these transcripts for the assigned stage categories. In addition to serving as extra *in silico* validation of the candidates, this approach may also provide a convincing method in molecular stage quantification for future studies. The primary validation testing was done by means of a 10 k-fold cross-validation testing (Figure 3.12) which serves as internal control (training datasets) of the algorithm's performance.

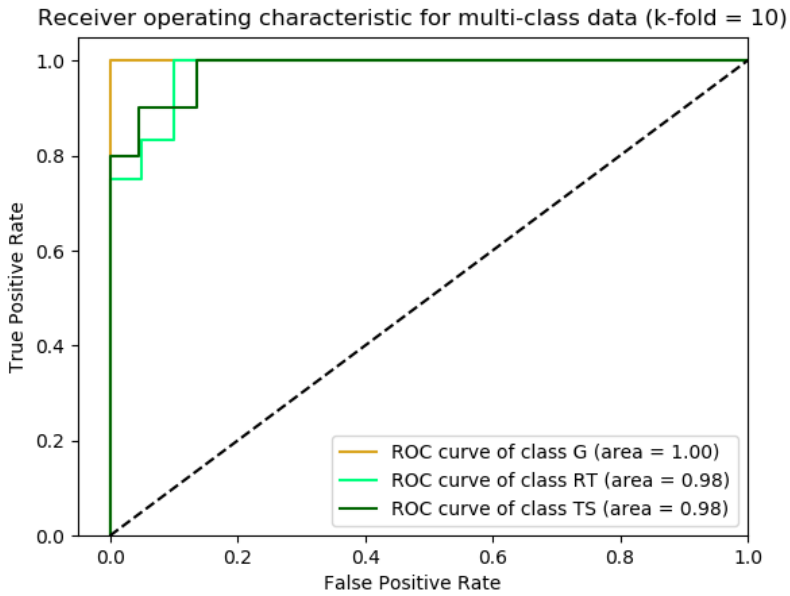


Figure 3.12: ROC output for stage-associated gene panel from *GradientBoostingClassifier*. True positive rate plotted over false positive rate with individual stage accuracy indicated. Colour legend used to indicate stage categories in gold (gametocyte = G), light green (ring-trophozoite = RT) and dark green (trophozoite-schizont = TS).

The best indicator of the model performance is done through test set evaluation. The accuracy achieved on the test set was 93%. To further confirm the validity of these genes candidates as a stage-associated panel, microarray data were integrated and assessed in the same manner (datasets from [48,95]), with accuracy achieved at 90%. Normalization strategies between microarray and RNA-seq data often produce several hurdles. For machine learning strategies the data distributions are the first hurdle to overcome, this was achieved through a simple mean centered normalization (forcing the mean to 0) and distribution ranges fell within margin between sets. The second hurdle is preservation of data order, given that microarray is a measure of relative abundance and RNA-seq is a measure of captured abundance, the data order is not preserved in the same manner between sets. A conversion of data based on extreme features such that the data is divided into 3 values (1, 0, -1) based on upper, middle and lower quantile thresholds should preserve data order for truly strong features in each condition set. Classifier model used: *scikit-learn GradientBoostingClassifier* with 5000 estimates.

The consideration of module assignment and stage-association of genes in this dataset are important features highlighted thus far and will be recalled throughout the following sections to dissect the transcriptome of gametocytes in more detail. Understanding of when transcripts are expressed and in what capacity will be fully utilized to interpret genes extracted via text-mining for regulatory potential.

3.3.6 Subnetwork construction of gametocyte related modules, highlight potentially key gene regulatory elements required in gametocyte maturation

Identifying potential regulatory factors for gametocyte maturation relies on all available gene ontology information from various sources that, in conjunction with the co-expression network and stage-associated genes, could resolve more important factors for this phase of the parasite life cycle. Text mining was used to filter genes which may form part of potential regulatory machinery in the parasite, and this was combined with specific modules relevant to gametocyte development. This subnetwork was filtered for the top 1000 interactions in modules 2, 3, 6 and 7 (henceforth referred to as sexual-associated modules) to gain an understanding of potential regulatory elements in gametocyte stages (Figure 3.13A). Modules 2, 6 and 7 show moderate correlation with gametocyte stages and are considered in these analyses, though module 2 and 6 correlated similarly to the TS category as well (Figure 3.13B). This is further evident by the average expression of these modules during gametocyte stages (Figure 3.13A). Elements with putative DNA-binding descriptions such as ApiAP2, zinc finger and CCAAAT-domain containing proteins were highlighted in this network.

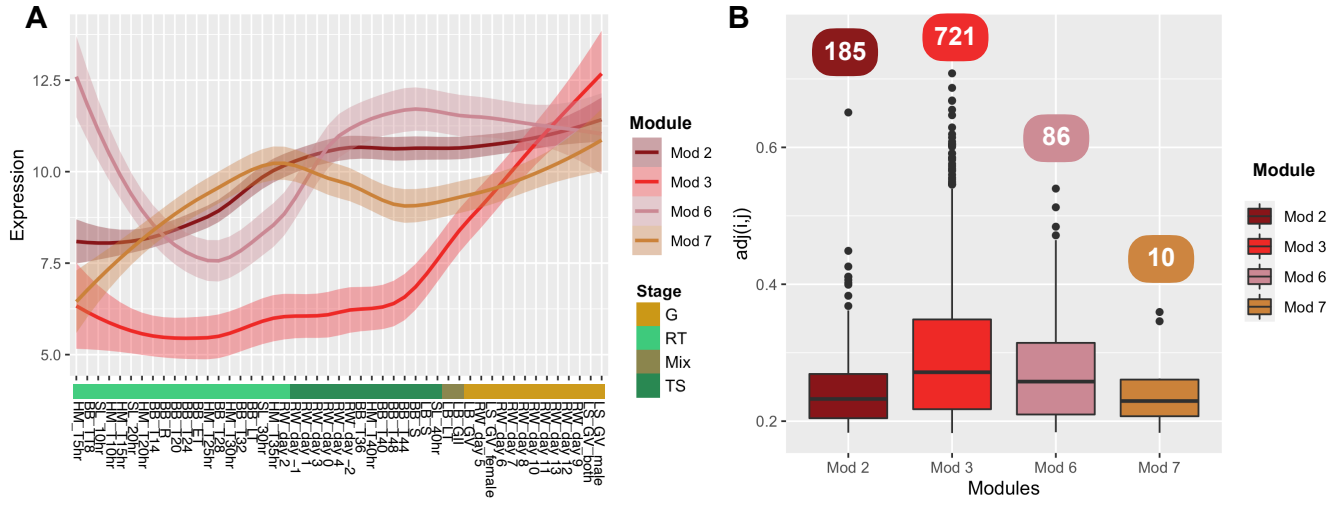


Figure 3.13: Subnetwork constructed from gametocyte associated modules and text-mining. A) Expression of modules 2, 3, 6 and 7 used in the gametocyte associated subgraph. Module expression shown over stage clearly illustrate modules involved in gametocyte stages. B) Module weighted correlation strength distributions and number of connections found in the subnetwork. Module connections are indicated in text boxes and colour coded.

Module 3 produces the largest community within this subnetwork (Figure 3.13B), producing highly connected hub genes that are very specifically co-expressed with sexual-stage specific transcripts, as this module is the best associated with sexual stage development. The entire subnetwork with genes to be discussed are labelled in the network topology and an associated

statistical breakdown table is also available showing the number of connections per gene discussed (Figure 3.14).

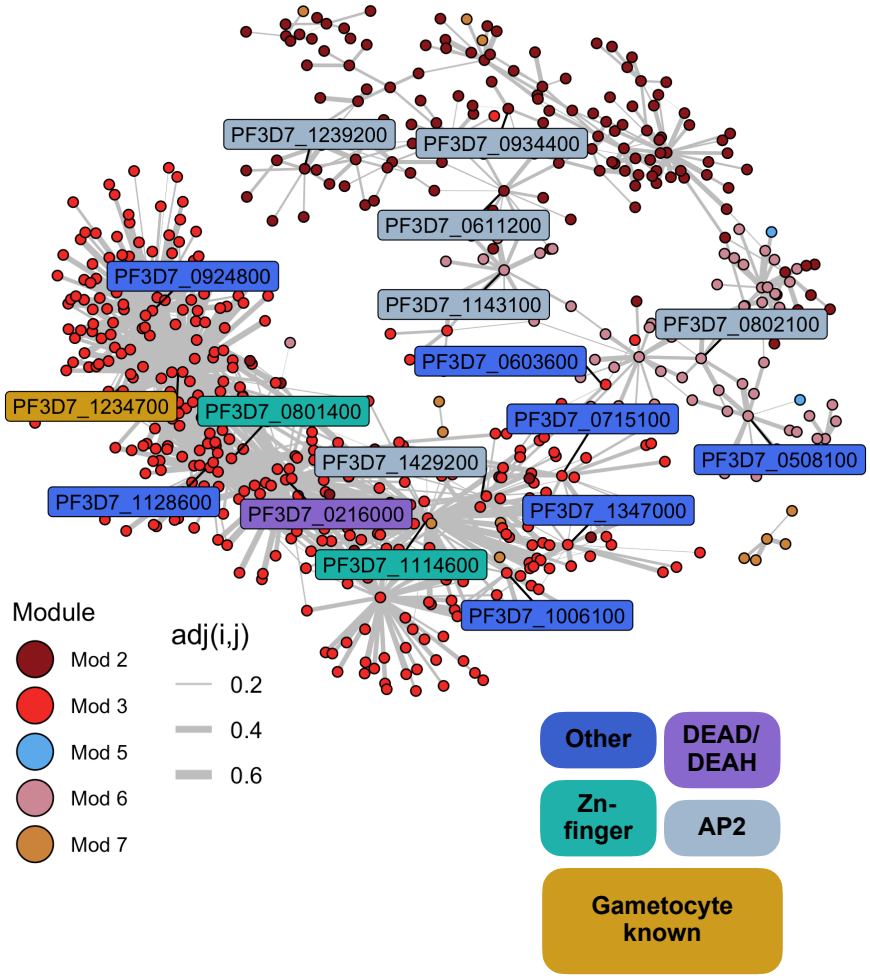


Figure 3.14: Subnetwork from gametocyte-associated modules.
Subnetwork graph with all modules present in the network. Genes discussed for their potential regulatory implications are labelled and position in the network is shown. Colour coding for Other, DEAD/DEAH, Zn-finger, ApiAP2 and known gametocyte genes are indicated.

Table 3.2: Associated statistical breakdown showing the number of connections per gene involved in the subnetwork in Figure 3.15. The associated breakdown of gene properties (in order of discussion) indicates gene id's, names, domain/functional, number of connections, sex-specificity for transcripts (Trx), proteins (Prot) or both (Trx + Prot), and translational repression (Rep). Male are indicated ♂ and female with ♀.

GeneID	Name	Domain Function	Connections	Trx	Prot	Trx + Prot	Rep
PF3D7_1234700	ULG8	CPW-WPC	150	120 ♀ 1 ♂	15 ♀ 8 ♂	122 ♀ 8 ♂	104
PF3D7_1006100	CCR4-NOT5	IPR007282	10	7 ♂	7 ♂	9 ♂	-
PF3D7_1128600	CCR4-NOT2	IPR007282	78	34 ♀ 8 ♂	8 ♀ 12 ♂	38 ♀ 14 ♂	22
PF3D7_0924800	TRF1	Homeobox-domain IPR001650	120	111 ♀	9 ♀ 1 ♂	111 ♀ 1 ♂	102
PF3D7_0216000	DEAD/DEAH	Helicase C	94	37 ♀ 24 ♂	9 ♀ 27 ♂	41 ♀ 31 ♂	28
PF3D7_1114600	Unknown	DNA-binding (GO:0003677)	100	6 ♀ 56 ♂	5 ♀ 53 ♂	10 ♀ 70 ♂	4
PF3D7_0801400	Unknown	DNA binding (GO:0003677) Transferase (GO:0016740)	48	14 ♀ 4 ♂	6 ♀ 7 ♂	16 ♀ 7 ♂	8
PF3D7_0802100	AP2	AP2-domain	10	1 ♂	2 ♀ 1 ♂	2 ♀ 2 ♂	-
PF3D7_0508100	SET9	SET-domain	15	1 ♀	1 ♀ 2 ♂	2 ♀ 2 ♂	-
PF3D7_1239200	AP2	AP2-domain	8	1 ♂	1 ♀ 1 ♂	1 ♀ 2 ♂	-
PF3D7_1143100	AP2-O	AP2-domain	12	3 ♀	-	3 ♀	3
PF3D7_0934400	AP2	AP2-domain	4	2 ♀	1 ♀	2 ♀	1
PF3D7_0611200	AP2	AP2-domain	14	3 ♀	2 ♀	4 ♀	1
PF3D7_1429200	AP2-O3	AP2-domain	21	10 ♂	5 ♀ 10 ♂	5 ♀ 13 ♂	5
PF3D7_0603600	ARID	ARID-domain	3	1 ♂	1 ♀ 2 ♂	1 ♀ 2 ♂	-
PF3D7_0715100	MYCBP	Regulation of transcription, DNA-templated (GO:0006355)	14	1 ♀ 7 ♂	9 ♂	1 ♀ 10 ♂	1
PF3D7_1347000	WD-repeat	WD domain (IPR001680)	17	7 ♂	9 ♂	12 ♂	-

A CPW-WPC protein, ULG8 (*pf3d7_1234700*) is highly transcribed in late-stage gametocytes and is highly connected within this community (150 connections). While the CPW-WPC proteins are translationally repressed in gametocytes[22,169], this transcript is closely correlated to seven of the other eight members of the CPW-WPC proteins and is closely co-expressed almost exclusively with other female-enriched or translationally repressed transcripts (127/150, 122/150 female enriched transcripts). This same overrepresentation of female-specific and translationally repressed genes was observed in the targets of CCR4-NOT

transcription complex subunits: *pf3d7_1128600* (*ccr4-not2*) and *pf3d7_1006100* (*ccr4-not5*), with 34/78 female-enriched transcripts (58/78 translationally repressed) in targets of NOT2 and 7/10 male-enriched, 3/10 translationally repressed targets of NOT5.

Telomeric repeat binding factor 1 (*trf1*: *pf3d7_0924800*) nested within the module 3 subnetwork presents with high connectivity and centralised topology in the network and shows the highest overrepresentation of female and translationally repressed transcripts in the network (111/120 female, 110/120 translationally repressed). TRF1contains a homeobox-like domain (IPR009057) and belongs to a family of proteins that have wide DNA-binding activities and are involved in e.g. recombination through to transcriptional regulation. It is not surprising that this cluster which so closely relates to a late-stage gametocyte transcriptional signature is enriched for female-specific genes, given that female-enriched transcripts peak later in gametocytes than male-enriched transcripts[95].

The trend for highly connected nodes in module 3 to be female enriched is broken by one specific putative regulator, a DEAD/DEAH helicase (*pf3d7_0216000*) which co-expresses with both female (41/94) and male (31/94) enriched transcripts respectively. DEAD/DEAH helicases can act as important cofactors to aid coactivation or co-repression of specific transcription factors and are themselves usually highly regulated (InterPro: IPR001650), which makes its expression at this stage of development noteworthy and is also interesting given that *pf3d7_0216000* is itself a male-enriched transcript. Thus, not all genetic regulation enacted in late-stage gametocytes is binarily linked to either male or female gametocytes but can impact the expression of sexual-stage specific genes. Further, an uncharacterised protein pf3d7_1114600 with a predicted gene ontology related to DNA binding forms a central node in the network is completely uncharacterized but shows a strong correlation with male enriched transcripts (70/100). These results suggest that although male-specific transcripts peak early in gametocyte development[95], a male-enriched transcriptional signature is distinctive in late-stage gametocytes.

The final highly connected node in this cluster, uncharacterised protein pf3d7_0801400 also with a predicted gene ontology related to DNA binding, forms one of the most central nodes within the subnetwork and similarly to the DEAD/DEAH helicase, is not highly connected to male (7/48) or female specific (16/48) proteins or transcripts. Together, this cluster highlights several under characterized genes that closely associated with sex-specific and gametocyte-specific genes that are of relevance for further characterization.

3.3.8 Putative specific transcriptional regulators co-cluster in separate early/late expression clusters

Genetic knock-out studies of the entire ApiAP2 family of transcription factors have been instrumental for the phenotypic and functional characterization of these proteins in rodent malaria models[134]. However, parallel studies in *P. falciparum* are hampered by its low transfection efficiency and haploid genome[170]. We aimed to predict which putative specific transcription factors could impact gametocyte development in the *P. falciparum* parasite. Within large transcriptional networks, ApiAP2 proteins are often under-represented given their subtle increases in transcriptional abundance and sometimes indirect effects on actual transcriptional abundance of their experimentally determined target genes[41,154,171] and reliance on additional regulatory factors, like the epigenetic regulators bromodomain protein 1 (BDP1) and HP1. However, such analyses do inform on those transcription factors that can show strong, direct regulation of some of their target genes ex AP2-G[38]. Within this gametocyte-selective subnetwork, we find most of these same putative regulators again, *pf3d7_0934400, pf3d7_0611200, pf3d7_0516800.* Interestingly, these are all associated with module 2, which skews towards gametocytes but also shows moderate correlation to TS stages. The one notable absent genetic regulator, *pf3d7_1222600 (ap2-g)* is expressed in a very tight window immediately preceding gametocyte development[38,44] and only within a subset of parasites committing to gametocytes, and these factors are expected to have confounded clear association with regulated genes in this network.

A small subcluster centred around module 6 included an ApiAP2 (*pf3d7_0802100*) and histone-lysine N-methyltransferase, *set9*, which were of interest as this module correlates with gametocyte genes but is slightly more correlated to TS stages. It is possible that these aforementioned genes play a minor role in gametocyte stages but interestingly our previous work[94] showed these proteins might play a role in the cell cycle of asexually dividing parasites and also picks out 3 of the same genes regulated by *set9* in our GRENITS network, (*pf3d7_0806300, pf3d7_0704500, pf3d7_1141800*) a ferlin (involved in vesicle fusion), protein kinase and EELM2 domain containing protein respectively and once again confirming an indirect interaction with *pf3d7_0802100*. Similarly, two ApiAP2's in module 2 showed a direct association (*pf3d7_1239200* and *pf3d7_1456000*) but were also skewed towards TS stage genes.

Module 6 also contained *ap2-o* (*pf3d7_1143100*), a translationally repressed transcript which co-expressed with female transcripts (3/12) that were also translationally repressed but along

with *pf3d7_0934400* (2/4) and *pf3d7_0611200* (4/14) represented the only ApiAP2 proteins in the network with slight biases towards female-specific transcripts or proteins. In addition to these previously predicted regulators, the subnetwork also highlights a cluster of genes that were strongly associated with gametocyte-related genes (module 3) and identified as possible transcriptional regulators that are expressed early in gametocytogenesis in our previous work (Chapter 2)[95]. These genes include *ap2-o3* (*pf3d7_1429200*), an uncharacterized AT-rich interaction domain (ARID) containing protein (*pf3d7_0603600*) and c-Myc-binding protein (*mycbp, pf3d7_0715100*). Interestingly, the group these genes were clustered into in our previous work (Chapter 2)[95] was also enriched for male-specific transcripts and all 3 of these putative regulators co-expressed with genes enriched for male-specific transcripts (13/21 for *pf3d7_1429200*, 2/3 for *pf3d7_0603600*, 10/14 for *pf3d7_0715100*). The transcript for a WD-repeat protein (*pf3d7_1347000*) strongly associated to *mycbp*, highly connected in the sub-community and similarly co-expresses with male-enriched transcripts or proteins (12/17).

High connectivity to the module 3 cluster could be indicative of a central role in gametocyte biology and these highlighted candidates could be of particular interest to study for their roles in male gametocyte differentiation. Together this subnetwork forms a source for well-established as well as novel transcripts related to sexual stage development that could be mined to provide leads for investigating sex-specific and sexual-stage specific processes in *P. falciparum* gametocytes arising from a well-curated and controlled set of combined RNA-seq datasets.

3.3.9 Adjacent neighbour gene pairs show positive correlations which could indicate shared promoters

The weighted co-expression profiles further led us to examine the effects of neighbouring gene pairs in the transcriptome, with the potential deconvolution of shared promoter effects. Though we cannot resolve the overlapping effect of neighbour gene transcription, the result (transcript co-expression or anti-correlation) may produce valuable insights. It is therefore necessary to investigate potentially anti-correlated effects in the co-expression network to account for neighbour genes that may negatively affect each other. To this end a modification in the WGCNA calculation, called *csuWGCNA,* was introduced to capture these anti-correlated results.

Both negative and positive weighted correlations (*csuWGCNA*) for neighbour gene pairs were calculated and illustrated across their Pearson correlations to indicate direction (Figure 3.15A).

The *adj(i,j)* scores therefore indicate the strength of the interaction (scale-free topology requirements met), while the Pearson correlations indicate the direction of the interaction (correlated or anti-correlated). A clear co-expression feature is observed for neighbour gene pairs rather than an anti-correlation feature as is evident in the skew towards correlated (Figure 3.15A). Furthermore, given that anti-correlated pairs rarely exceed *adj(i,j)* above ~0.4 it seems unlikely that this network has captured any neighbour pair interference in expression. The network does however seem to indicate strong co-expression with values as high as ~0.8, which may be helpful in understating shared promotors.

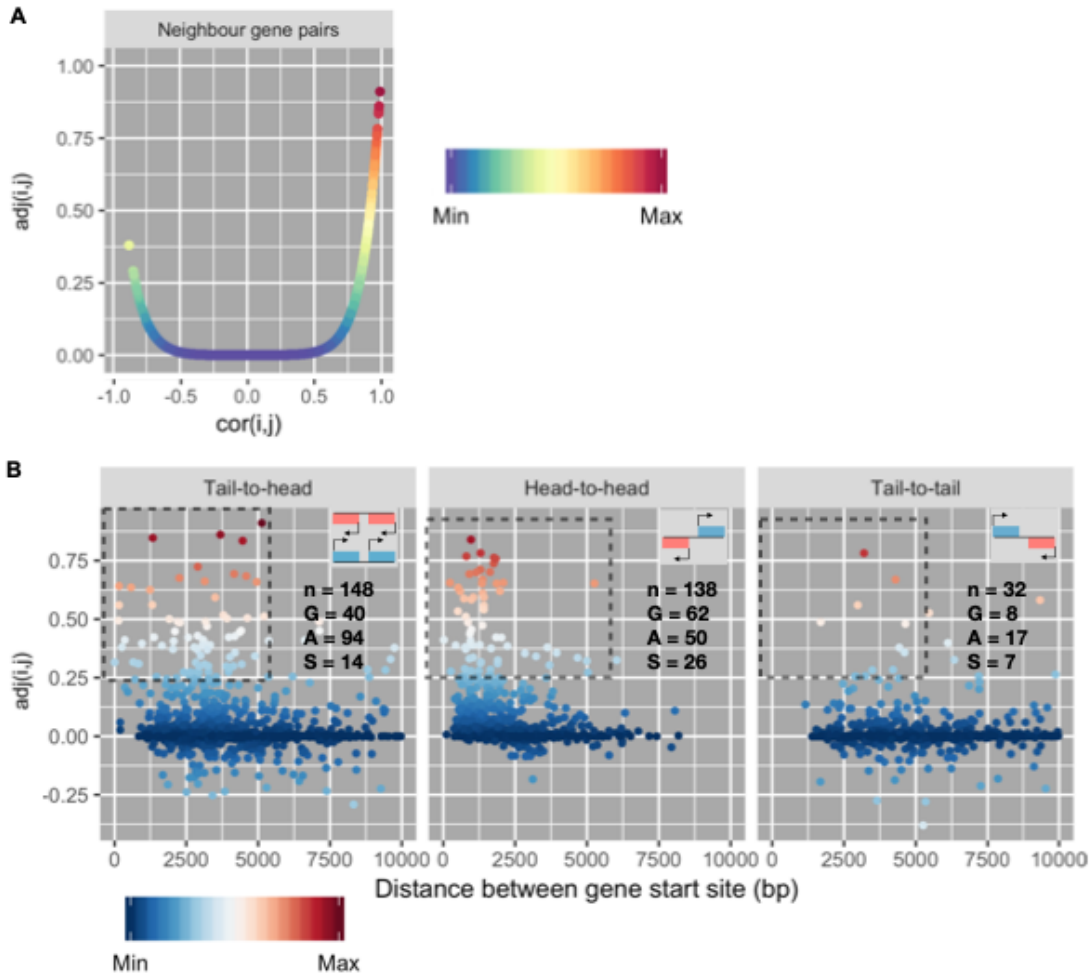


Figure 3.15: Neighbouring gene pairs captured using *csuWGCNA*.
A) Neighbour gene pairs and their weighted correlations as show with regards to negative or positively correlated relationships. *adj(i,j)* = *csuWGCNA* adjacency values, *cor(i,j)* = Pearson correlations. B) Neighbour gene pair weighted correlations for 3 configurations (Head-to-tail, head-to-head and tail-to-tail) shown over distance between gene body start sites. Striped boxes indicate gene pair associations *adj(i,j)* > 0.25 and a distance between gene start site < 5500 bp. Values of the striped boxes are show; n = total, G = number in gametocyte modules (3,7), A = number in asexual modules (1,4,5,8,9,10,11) and S = number of genes in shared modules (2,6).

Distance between genes is often relevant in shared promotor formations. Thus, gene neighbours were subsequently also arranged into three configurations that show gene

orientation with regards to their adjacent neighbour gene pairs: tail-to-head (n = 2401), head-to-head (n = 1678) and tail-to-tail (n = 1702) configurations (Figure 3.15B). Neighbouring gene pairs show strong co-expression signals suggesting that they are either likely to have shared promotors and potentially shared transcriptional factors/complexes with separate promotors or alternatively co-express independently of shared regulatory mechanisms. However, the expression of a neighbouring gene does not seem likely to be repressive regardless of configuration. Head-to-head and tail-to-head configurations are the most prominent, with very few tail-to-tail associations found. This would be expected as the distance between tail-to-tail configurations may be vast. This configuration in the linear plane ignores the potential three-dimensional effects of the genome which may bring some gene pairs in closer proximity.

Head-to-head configurations are most indicative of potential bi-directional promotors and stronger co-expression is observed for those in closer proximity of each other (Figure 3.15B). Many of these pairs have been observed before[24].

Neighbouring gene pairs were further subdivided into module association along strict lines to gain insight into which phases are relevant (asexual, gametocyte or shared between the two). Modules which only associated with gametocytes were used to count the number of gametocyte neighbours which co-express (Figure 3.15B), the same for asexual modules. Shared modules which present with moderate correlation to both asexual and gametocyte stages was similarly used. Tail-to-head configurations would appear to be largely relevant for asexual co-expression (94/148), which possibly indicates the use of tandem expression. Similarly tail-to-tail configuration appears to involve asexual co-expression (17/32), however these numbers are relatively small. More gametocyte related neighbour co-expression in the head-to-tail configuration (62/138), however many asexual neighbours also co-express (50/138). The differences between gametocyte and asexual co-expression for these neighbouring configurations are small, and perhaps expected. It seems more likely that neighbouring genes are influenced by the gene orientation (which provides us with the above configurations) and access to the genome through the chromatin structure. Transcription start sites may also influence neighbour gene co-expression, as was noted for *pf3d7_0505500*/ *pf3d7_0505600* (head-to-tail), the closer the TSS of *pf3d7_0505500* was to *pf3d7_0505600*, the more pronounced their co-expression[24]. Those corresponds to the fact that head-to-tail pairs which are closer to one another have stronger co-expression. The relevant TSS for gametocytes are however not known for *P. falciparum*. It is therefore more plausible that these configurations and their co-expression are driven by localised factors such as gene orientation and chromatin structures, and less likely due to differences in parasite phases.

3.3.10 Intergenic lncRNA show both strong co- and anti-correlation with genes:

As previously discussed, the role of intergenic lncRNAs such as lncRNA-TARE have been shown to play a role in gene regulation. The mechanisms with which in lncRNAs can regulation transcripts is still unclear, however elucidating potential target transcripts may be a step closer to understanding these mechanisms. Newly discovered lncRNA were defined as transcripts which do not reside within nuclear gene bodies and had a length greater than 100bp. These newly discovered lncRNA's are named based on the chromosome they are found and roughly the start to end bp numbers (pf3d7_chromosome_v3: start coordinates – end coordinates). Intergenic lncRNA's (new and existing) were correlated against the transcripts from nuclear genes with *csuWGCNA* to evaluate both positive and negative weighted co-expression (Figure 3.16). Strong co- and anti-correlation is observed for these lncRNA with that of nuclear genes (Figure 3.16). For the pairs that are anti-correlated, the explanation may be as simple as occurring during different stages and thus have no significant bearing on each other. For instance, if a transcript expresses during asexual development and the anti-correlated lncRNA is only present during sexual stages, this could just be a coincidence. However, given that the mechanism of how lncRNA's can affect transcription is not entirely clear the presence of this lncRNA might be required to repress the aforementioned hypothetical transcript. Such causal relationships cannot be inferred from these data, thus we only take note of these correlations. Correlated pairs on the other hand may be a good indication of post-transcriptional repression, however as with anti-correlated pairs we can at most note these relationships, as inferring causal relationships will require experimental validation.

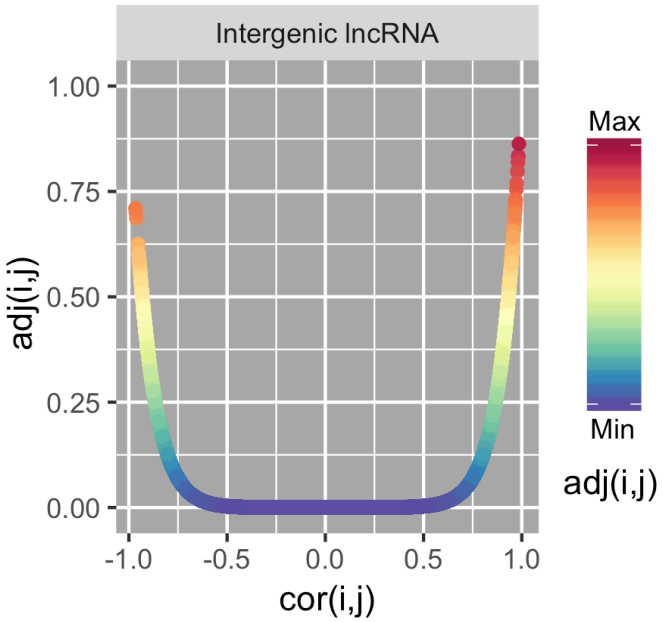


Figure 3.16: Intergenic lncRNA captured using *csuWGCNA*.
Intergenic lncRNA and weighted correlations to nuclear genes as visualized. Interaction strengths *adj(i,j)* and Pearson correlations *cor(i,j)*.

A more in-depth analysis of these relationships required separating noisy correlations from non-noisy ones. Intergenic lncRNA's were therefore assessed for significance in terms of their association to nuclear genes using generalized linear modelling in attempt to go beyond simple correlations. A stringent cut-off was imposed for these associations with adjusted *P*-value < 0.001 and weighted correlations greater than 0.4. These significant relationships were used to construct a subnetwork of the data. The subnetwork was further divided into gametocyte correlated and gametocyte anti-correlated subsets to investigate potentially interesting lncRNAs (Figure 3.17).

Four noteworthy lncRNAs correlated with genes associated to gametocyte stages: *pf3d7_0108900, pf3d7_0829700, pf3d7_0935390* and *pf3d7_0809300* (Figure 3.17A). Comparison to translational repression and transcript stabilisation was used to infer if these lncRNA play a role in repression of their respective target transcripts. Male repressed transcripts were observed for *pf3d7_0108900* (n=3 repressed, n=6 stabilised) and *pf3d7_0935390* (n=1 repressed), conversely *pf3d7_0829700* correlated with 4 translationally repressed female transcripts, while *pf3d7_0809300* did not correlate with any repressed transcripts.

The subsets of lncRNAs which anti-correlated with gametocyte transcripts are *pf3d7_1229200, pf3d7_0626840, pf3d7_02_v3:409523-409680(+)* and *pf3d7_0918500* (Figure 3.17B). Two lncRNAs anti-correlate with a large number of gametocyte-associated

genes, *pf3d7_1229200* (64/74) and *pf3d7_02_v3:409523-409680(+)* (73/117), with the other at relatively low proportions of the overall connections, *pf3d7_0626840 (35/132)* and *pf3d7_0918500* (2/9). Both *pf3d7_1229200* and *pf3d7_02_v3:409523-409680(+)* have a strong inverse relationship with gametocyte genes with peak expression during early asexual stages (RT). Previously, mechanisms involving lncRNA associated at telomere regions were postulated to repress the transcripts of a *var* gene[61]. Though much remains unclear about how lncRNA can regulate the parasite transcription, it is tempting to investigate the potential role of *pf3d7_1229200* and *pf3d7_02_v3:409523-409680(+)* in repression of their respective gametocyte target transcripts.

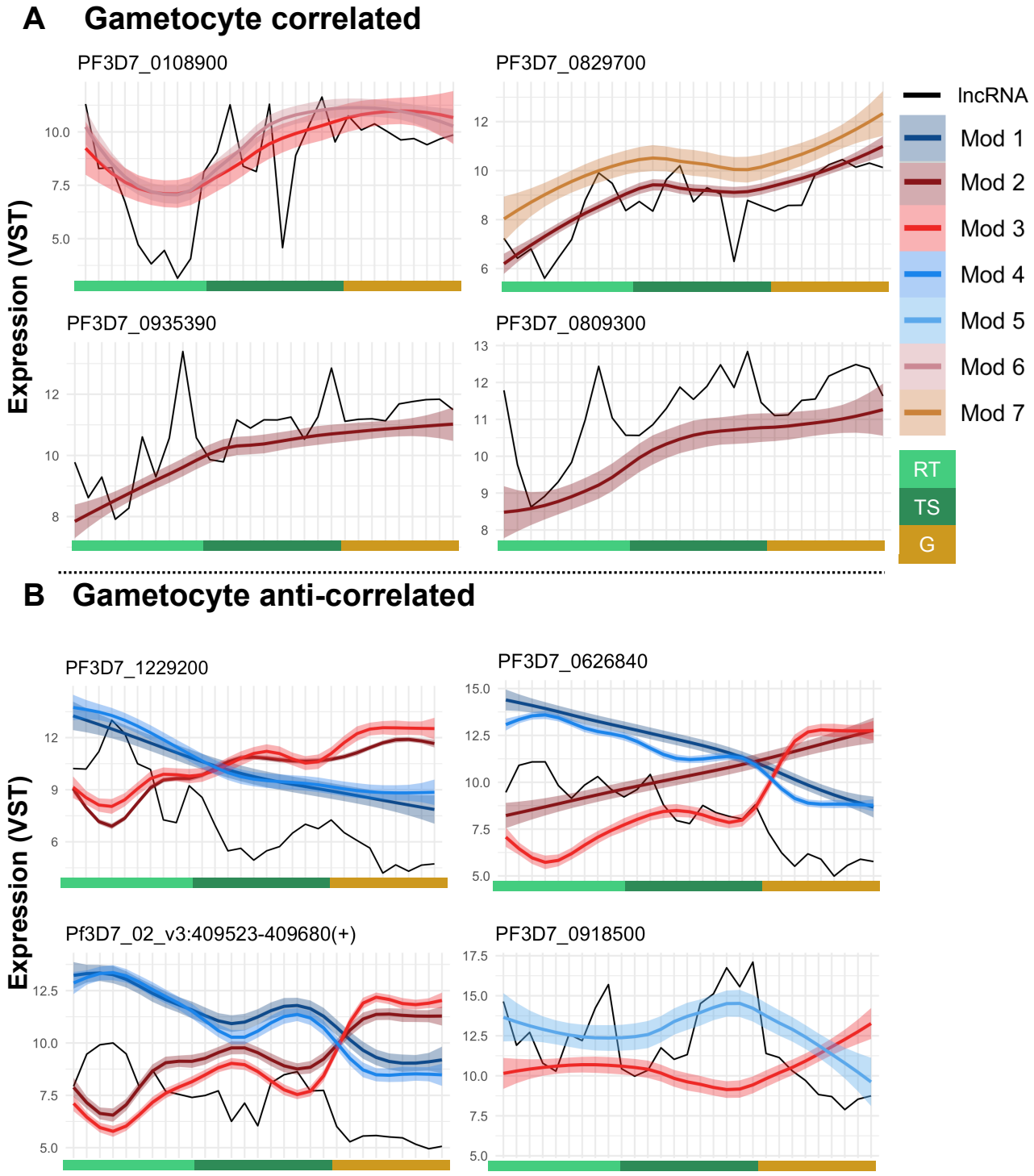


Figure 3.17: Subnetwork of lncRNAs show correlated and anti-correlated expression with gametocyte modules.
Subnetwork of lncRNA and interacting nuclear genes filtered for adjusted P-value < 0.001 and a weighted correlation > 0.4. A) Subset of lncRNA which correlate with gametocyte modules. Expression of lncRNA shown in black and modules coloured. B) Subset of lncRNA which anti-correlate with gametocyte modules. Potentially novel lncRNA's discovered during the analysis was annotated as strain of genome followed by version and then the start and stop coordinates (pf3d7_chromosome_v3: start coordinates – end coordinates).

3.3.11 Intragenic lncRNA (anti-sense transcripts asRNA) potentially repress transcripts during gametocyte stages

Previously, asRNA mechanisms of transcriptional regulation have been characterised and shown to play roles in regulation, such as *gdv1* asRNA, however little is known about asRNA regulation in gametocytes. We therefore evaluated the relative abundance of anti-sense transcripts, as compared to their corresponding mRNA transcripts with the DAFT-seq time course data. These asRNAs were quantified as FPKM-normalized reads of their corresponding gene body regions and a log_2FC ratio (asRNA/mRNA) that was calculated followed by k-means clustering (nine clusters, Figure 3.18). Clusters 4 and 5 are relevant to gametocyte stages as they present with high asRNA:mRNA ratios during these stages. In cluster 4, we find chloroquine resistance transporter (*pf3d7_0709000*) and drug/metabolite transporter *dmt2* (*pf3d7_0716900*), which are required during asexual development and has not been assigned a specific function during sexual development stages (PlasmoDB cross referenced for essentiality). Other genes such as *etramps* are also seen here, indicating that these strongly asexually associated genes are repressed in sexual development and that this repression may involve asRNA. Cluster 6 and 9 show strong asRNA signal throughout the time course, suggesting asRNA transcripts are present at higher abundance for these clusters throughout the time course. Gene ontology analysis of these clusters suggest they are mainly involved in transport processes and protein lipidation. Principle amongst genes in cluster 6 is and essential protein involved in sexual commitment steps, *gdv1* (gametocyte development protein 1, *pf3d7_0935400*), which is known to be repressed by its asRNA transcript[45]. Four conserved unknown genes occur in cluster 6 (*pf3d7_0902900, pf3d7_0934600, pf3d7_1005800* and *pf3d7_1318000*). These four genes have no discernible function in the parasite apart from InterPro domain results that indicate the presence of membrane components and transporter domains. Two rifins also feature in this set, asRNA would appear to repress their sense transcripts, however assessing rifins or virulence genes from laboratory-based cultures do not always produce sound insights as the same environmental pressures are not replicated i.e. immune system of the host.

In cluster 1 and 8 there are transcripts with high levels of asRNA signal during asexual stages and low signal in gametocyte stages. Cluster 1contains sexual stage-specific protein G37 (*pf3d7_1204400*), suggesting these transcripts are functionally more important in gametocyte stages but might need to be repressed during asexual development. Such mechanisms have been observed for example the repression of *gdv1* during the IDC where asRNA plays the key role in preventing the initiation of gametocytogenesis. Singh et al.,[41] also found sets of gene

repressed during the IDC (active in sexual development) which were not linked to repressive transcription factors, which would imply alternative mechanisms perhaps such as asRNA. A conserved unknown gene (*pf3d7_1302400*) in this cluster contains transmembrane helices and may be an integral membrane component of gametocyte stages as it was found to be specific for expression in this stage. In cluster 8 (n = 11), *pf3d7_0516500* was found to have strong asRNA signal during asexual stages and low signal in gametocyte stages. *pf3d7_0516500*, a transcript for a major facilitator superfamily domain-containing protein, may play a role in transmembrane transport and was also shown to be specific for gametocyte stages. Another potential transmembrane protein (*pf3d7_0912200*) is also found in this cluster with gametocyte specificity. Neither asRNA transcripts seems to be explained by their adjacent gene pair expression i.e. overlapping UTRs from neighbouring genes was not observed which could have been counted as asRNA. The independence of these asRNA findings was cross-referenced with neighbouring genes to ensure they are not the product of neighbouring gene UTR overlaps. This may suggest independent transcription of these asRNA.

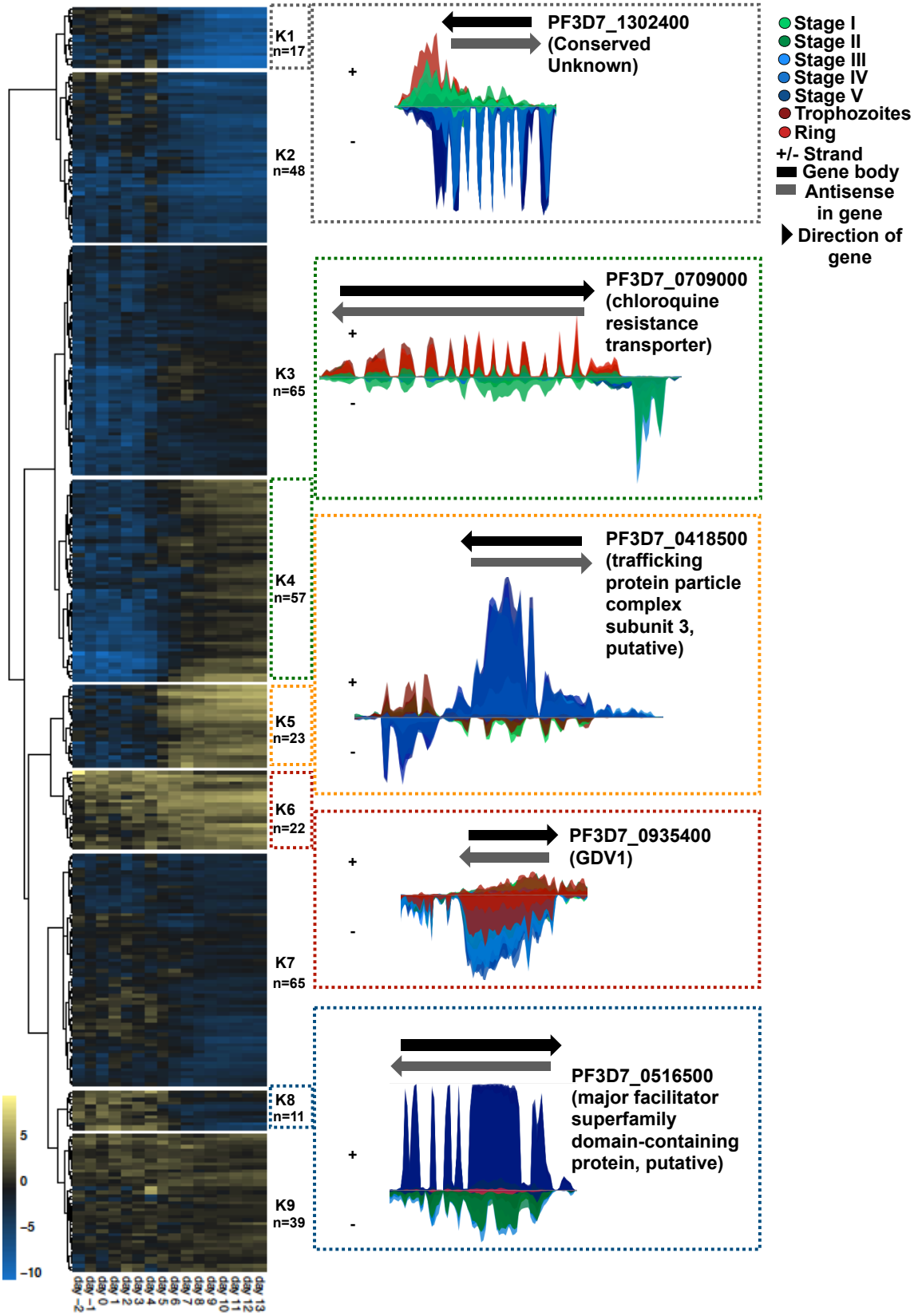


Figure 3.18: Intragenic anti-sense RNA (asRNA) for in-house time course dataset (day -2 to 13). The asRNA transcripts were calculated as a log_2FC(asRNA/mRNA) ratio with gold indicating high asRNA abundance. K indicates clusters and their respective numbers. Stage I-V refer to gametocyte stages and colours match genome coverage plots. Sense and antisense directions are indicated for each example gene. Stands are indicated by (+/-) denoting positive and negative strands respectively.

3.4 Discussion

This study combines data from key *Plasmodium* RNA-seq datasets to build out strongly stage-associated transcriptional markers through powerful network analysis approaches and highlights the importance of asRNA and lncRNA in regulation of gene expression in *P. falciparum* sexual development. Many factors driving gametocyte development remain poorly understood, here we quantify the relationships between genes and their relative stage associations in the parasite, allowing us to further investigate potential factors which affect transcription in gametocytes.

This study shows the construction of comprehensive co-expression network contrasting different parasitic life phases, which accentuates the key contributing factors for regulation during each phase and their subsequent stages. The power of this analysis comes from the use of effective cross-normalisation between the datasets. This variable stabilising transformation escaped the trappings and caveats normally encountered by combining different datasets and the constraints and biases introduced by different experimental datasets. This allowed us to compile a strong co-expression network of stage-associated transcripts, highlighting a clear progression and separation of early asexual development from late asexual development carried through to sexual development. Prominent gene clusters are formative in the network topology and describe the split nature between the two life phases. Intramodular hub genes show high connectivity and strong correlation to their respective modules, thus are core features of these modules. This would suggest that these hub genes are needed for their respective modules and inferred development stages.

The power of correlated associations highlights the importance of the hub genes at the core of their respective development stages, the rationale being that if these genes are strongly associated with other genes in this development cycle and that their presences are unlikely to be coincidental. We propose this panel of genes as informative for predicting parasite stages in terms of the set categories from transcript levels in conjunction to the stage associated gene modelling that was performed. The vast majority of RNA-seq datasets for *P. falciparum* have focused on asexual development and here, we contribute a full gametocyte maturation time course RNA-seq dataset to produce a more balanced network that can capture the total complement of stages in asexual and sexual blood stage development of the parasite. Except for the construction of a detailed co-expression network highlighting key genes for stage-transition in *P. falciparum* parasites, our in-house RNA-seq dataset also enabled the exploration of additional transcriptional control mechanisms (asRNA and lncRNA) that could not be investigated in previous microarray time courses[95,105].

Beyond the interplay of genes and their effect on one another, the role of alternative regulatory mechanisms such as non-coding RNA and the localised effects of neighbouring gene pairs also show possible involvement in gametocyte development. Neighbouring gene pair correlations proved overwhelmingly in favour of positive relationships suggesting the potential for shared promotors, though the granularity of the data may produce asynchronous artifacts which the co-expressions could not account for. However, for the pairs where these relationships do occur, it would be interesting to further investigate the potential role of a shared promotor in these regions.

Other lncRNA occurring outside of gene body such as intergenic lncRNA may also play a role in gene regulation. It was previously proposed that some lncRNA, particularly those situated near telomeres have a regulatory role for *var* genes[155]. lncRNA have been showed to play a role in gene regulation in other organisms, particularly for Apicomplexa[42]. However, no such relationships have been investigated for gametocytes in *P. falciparum*. Here, we identified novel gametocyte specific lncRNA with strong co-expression and anti-correlated relationships with gametocyte-related genes. The effect these lncRNA have on the expression of genes remains to be seen, but their existence and quantified relationships do present interesting research prospects. Validation of lncRNA and their respective targets is a complex problem and most likely require knockout or knockdown validation studies to measure their effect.

The power of this study resides in tailoring the approach to the data at hand. Noisy relationships between gene pairs are par for the course for any researcher dealing with large transcriptomes spanning multiple life phases, sometimes generated at different times by different researchers and in different labs. Usually, the addition of more data can introduce more noise, however, with this approach, we managed to cut to the core of potentially meaningful relationships between data points. This expedites the data discovery process and helps to gain more insights into the underlying data structure of relevant factors. These data allow for the probing of more complex questions and more extensive analysis, modelling the predictive power of relevant regulatory candidates (Chapter 4).

Chapter 4

The application of Gene Regulatory Networks in *P. falciparum* through inference-based machine learning approaches.

4.1 Introduction

In chapter 3 we constructed a large unsupervised network through co-expression techniques, however, co-expression datasets tend to lack predictive power. In chapter 2 we produced GRNs with high predictive power, but were constrained by the scale of the data. Therefore our next aim was to employ an algorithm which could both expand the scale of the analysis as well as the predictive power of the relationships. Decision tree algorithms such as RF and GBM allow for expedited assessment of many genes compared to DBNs. RF algorithms such as GENIE3 have gained a lot of popularity and are considered a reliable approach to GRN construction. This process makes use of supervised learning strategies whereby the core model consists of candidate genes and training occurs against target genes in an iterative manner[71]. The importance inferred from these training-set-per-target-gene forms the basis of the candidate-target interactions. The inferred importance is effectively a ranking of the feature relevance or contribution in the prediction of the target gene's expression during training and identifies which of the candidate genes were most informative in predicting the outcome. The individual importance also strongly correlates with the accuracy of the predictions[71,78,79].

A key drawback of RFs, however, remains a resource issue: to obtain accurate results, many trees need to be constructed as the consensus is what makes RFs so accurate. This is where GBMs have the advantage as they learn from each tree building step (shallow decision trees) rather than consolidating completely constructed trees as is the case with RFs. One attribute of GBMs reigns supreme, that is they are easy to train and generally considered reliable. A prominent GBM developed with this goal in mind, GRNBoost, that references GRN gradient boosting, was developed by the Laboratory of Computational Biology (Aertslab: https://aertslab.org)[79]. This first version required XGBoost at the core of its engine but has since been updated to include Dask as a parallel computing library in conjunction with Scikit-learn GBMs (GRNBoost2). This was incorporated into a full analytics suite called Arboreto[78]. GRNBoost2 is considered to be Random Access Memory (RAM) intensive, however it depends entirely on the size of the submitted data. The resource requirements for GRNBoost2 vary depending on the gene number and sample size included in the analysis. The original platform served to analyse human single cell RNA-seq (scRNA-seq) with thousands of

samples. *Plasmodium falciparum* has a much smaller genome and far fewer genes to evaluate, and the resources required to perform these analyses are generally less intense.

Here, we investigate the usefulness of GRN analysis in the *P. falciparum* research field and its ability to infer relevant regulatory candidates. The investigation is divided in two parts. We start by constructing a global GRN, which reflect both asexual IDC and gametocyte phases of development resulting in the most comprehensive GRN in the field to date. Here, a wider "net" is cast on both candidate and target genes by employing the power of GBM based techniques such as GRNBoost2. Lastly, the construction of GRNs through Arboreto, was packaged in a "user-friendly" application for researchers in the field and provides access to a pre-compiled filterable GRN constructed during the study.

The content of this chapter has been presented in part in the following instance:

1. **van Wyk, R.,** van Biljon, R., Birkholtz, L. (2021) MALBoost: a web-based application for Gene Regulatory Network Analysis in *Plasmodium falciparum.* Malaria Journal. 317.

4.2 Methods:

4.2.1 Consolidate gene candidates for construction of a global GRN in *P. falciparum*

Gene candidates discovered through three studies (chapter 3, cell cycle[96] and ApiAp2's from the gametocyte network) were used as putative regulatory genes. These candidate genes were further filtered manually based on biological relevance for a total of 124 candidate genes. Biological relevance was as assigned to candidates which showed descriptions for DNA-binding and transcription regulation as based on known and predicted hits from GO and InterProScan. Known and predicted nuclear localisation was also used to inform the selection. Candidate genes were submitted to PlasmoDB.org for GO analysis. A threshold of *P*-value < 0.05 was used.

4.2.2 Global GRN for *P. falciparum* using GRNBoost2

All candidate genes identified through the previously described consolidation process (total of 124) were evaluated against 5163 target genes using the integrated RNA-seq dataset compiled in chapter 3, with VST values as described in the normalisation steps (Chapter 3). Candidate genes were used to construct and global GRN for *P. falciparum* using GRNBoost2 from the Arboreto suite. GRNBoost2 uses gradient boosting machines from scikit-learn in conjunction with Dask for task queuing and job prioritization[78]. This tool uses supervised regression learning to infer the relative importance of features (in this case candidate genes) for each target vector (target genes). Candidate genes are ranked for importance following the algorithm output and the top 100 set of target genes are submitted for motif discovery using DREME and FIRE[172,173]. Filtering for the top 100 interactions per candidate gene was done in accordance with the methodology of the original GRNBoost publication[79]. This ensures the evaluation of the strongest interactions found per candidate as the tool would have evaluated each candidate against all target genes and generated a score.

4.2.3 Motif discovery of candidate genes

The top 100 target genes following GRNBoost2 analysis for each candidate gene was used in two separate motif discovery tools. Genes were submitted to the FIRE online tool [174].

Upstream sequences of target genes from 1500 bp upstream of gene start site and 500 bp downstream of gene start site were used as input. Strandedness of genes were preserved and reverse compliments applied to preserve the correct strand orientation. These target sequences were also submitted to the DREME online motif discovery tool[175].

4.2.4 Web-based application for user friendly access to Gene Regulatory Networks, MALBoost

To make the power of GRNBoost2 more accessible to a wider audience of researchers in the malaria field (particularly those with a limited background in Python or bioinformatics/Computational biology), the tools were packages in a "user-friendly" web-based application. This provides access to GRNBoost2, GENIE3 as well as a pre-compiled network produced from this study which are downloadable. The application is named MALBoost, after GRNBoost with intended use in the malaria research field. The application is constructed through Python-Flask which is a microframework for web applications built in python. Flask handles requests via the web interface and distributes the request to various technologies. Flask requires a Web Server Gateway Interface (WSGI) to handle these requests more efficiently, here Green Unicorn (Gunicorn) provides this service. Since the requests made potentially involve the construction of GRNs through either GRNBoost2 or GENIE3, background processes are required to run the request and does not require the user to keep an open connection with the web front-end while waiting for results. Rather the networks are constructed in the background and the results are emailed as a downloadable link to the user. In order to facilitate these background process two components are required, 1) a task queuing server and 2) an in-memory data structure store or broker such as Redis or RabbitMQ. For this implement we've made use of Celery as a task queuing server system with Redis as our data store for passing request data onto the Celery workers. This not only ensure that the user submission runs in the background, but also provides for multiple submissions being made with a sensible task queuing server addressing the submission on a "first come, first serve" basis.

General data such as email address and network results are temporarily stored in an SQLite databases (DB) for retrieval. The results are emailed to the user through Flask's own internal mail service in conjunction with a set Gmail account for the application. The email contains the result download link, the email may also contain any error status which occurred during the network construction. Data stored in the DB are temporarily stored for 3 days and deleted via the Flask application schedular. This ensure that the DB does not get congested, and that

personal details of users and results are not stored long term. Submission data (transcriptomes submitted) are passed onto the task queuing servers which run the GRN models through the SQL DB, however the transcriptome data itself is immediately purged from the DB upon model completion. This is to ensure the privacy of the user and to not strain the storage resources of the system.

The front-end architecture of the application is web-based as mentioned, which gives users access to the tools without requiring any installations or hardware resources of their own. The web-application facilitates the front-end using three popular technologies: Hyper Text Mark-up Language (HTML), Cascading Style Sheets (CSS) and JavaScript. HTML is responsible for the overall architecture of the web page with CSS controlling the web page style such as font type/size, colours, and various aspects of the web aesthetics. The web front-end is based on a modified version of pre-constructed bootsrap4 template (https://startbootstrap.com/themes/) which offer free and open-source templates for the use in web applications. The application is currently deployed on the University of Pretoria Centre for Bioinformatics and Computational Biology servers and can be found at http://malboost.bi.up.ac.za.

4.3 Results

Candidate genes with potential regulatory function in gene expression in different stages of *P. falciparum* were collected and consolidated from three studies*:* the transcriptome of the asexual cell cycle study[94] (referred to as VBC), the microarray-based transcriptome of gametocyte maturation[95] (referred to as VBG, which primarily focused on the ApiAP2 transcription factor family) and the RNA-seq transcriptome study performed in Chapter 3 of this thesis (referred to as VWG). Consolidation of candidate genes identified in each of the three datasets produced several overlapping candidates; five ApiAP2 genes between all three datasets and 22 genes shared only between the gametocyte-associated datasets (VBG and VWG), with VBC and VWG sharing 19 additional candidates and five genes shared across all datasets. A total of 305 candidate genes were identified, with 88 unique genes in the VBC dataset and 171 in the VWG dataset (Figure 4.1A). A final list of 124 candidate genes were identified for GRN construction that all presented with strong GO terms (*P-value* < 1.6e-06, biological process level, Figure 4.1B) for regulation of biological process, regulation of cell cellular process, regulation of transcription, multiple macromolecule biosynthesis regulation and regulation of gene expression.

Based on these candidates, a comprehensive list of genes, which potentially regulates transcription in *P. falciparum,* was compiled and was evaluated through the construction of a global GRN using GRNBoost2.

Figure 4. 1: Consolidation of candidate genes for transcriptional regulation in *P. falciparum*. A) Venn diagram of candidate genes as produced from the findings of three studies: van Biljon gametocyte (VBG)[95], van Biljon cell cycle (VBC)[94] and van Wyk gametocyte (VWG) in chapter 3. From a total of 305 candidates only 124 were considered for GRN post-manually filtering for biological relevance. B) Gene Ontology analysis of candidate genes rank from highest counts per biological process to lowest. P-values indicated as -log10 for both colour and size. All P-values < 1.6e-06.

4.3.1 Global GRN constructed through supervised machine learning escapes the trappings of conventional correlations and size constraints of Bayesian networks

The 124 candidate genes were used to construct a global GRN using GRNBoost2 by evaluating the candidates against a total of 5163 target genes (transcripts captured and shared across add datasets). This evaluation was done using the integrated dataset produced in Chapter 3 as it is the most representative of both parasite life phases (asexual and sexual) between the three studies with appropriate normalisation between sets and samples. It also produced the largest sample set of the three studies at n=49. Normalisation between microarray and RNA-seq data could not be achieved here hence the focus on the integrated RNA-seq dataset from Chapter 3 for the construction of the GRN. As described previously (Chapter 1) and in [78], the candidate genes are used as model features (the X in machine learning terms) in order to predict the target gene (or the Y) for each gene. The relative importance of each candidate gene is then extracted and assigned as the interaction between candidate gene and target gene. This theoretically produces 640 212 interactions (124 x 5163), however the tool does remove interactions that do not produce predictable outcomes and resulted in a total of 636 732 interaction for the network. This network size vastly exceeds our previously constructed Bayesian networks through GRENITS in Chapter 2.

The GRN captures 324 020 negative or potentially "repressive" relationships with 312 712 positive, co-expressed relationships (as assessed through Pearson correlations) to assign direction of the interaction relationship (Figure 4.2A). The general trend between correlation and inferred importance seen through a Generalized Additive Model (GAM) fit shows a general associated trend between the two metrics, but not a convincingly strong relationship (Figure 4.2A with r^2= 0.495). Thus, generally, the better correlated the interaction, the more likely the case that it will have strong inferred importance. However, many strongly correlated interactions ($r^2 \geq |0.7|$) do not seem to produce high importance values (< 5) Figure 4.2B. This illustrates that the model had low predictive power for many of the interactions which are strongly correlated, questioning the reliability of the correlations themselves and validity of "guilt-by-association". Interactions that strongly correlate but fail to exhibit predictive potential when models aim to predict the outcome, are likely to be coincidental in nature.

Conversely, instances where the correlations ≤0.5 exhibited importance values > 40 showed high predictive power which was unexpected (Figure 4.2C). This may be the result of GBMs finding asynchronous patterns in the dataset, which ultimately seem to predict the outcome with some measure of reliability. The inferences would therefore be able to tease out more

relevant interactions than correlations. There are however candidate genes, for which correlation strength ($r^2 \geq 0.5$) indeed matches that of inferred importance (>60) (Figure 4.2D). Interestingly, the predictive power of GRNBoost2 seems to capture correlated and anti-correlate interactions relatively well.

Machine learning, which aims to predict the outcome more accurately, therefore escapes the trappings of conventional correlations and highlights that many of the interactions that would normally be heavily emphasised due to their high correlation strength, may lack relevance in biological interpretations. The data here therefore provides a useful list of important gene candidates as regulatory elements.

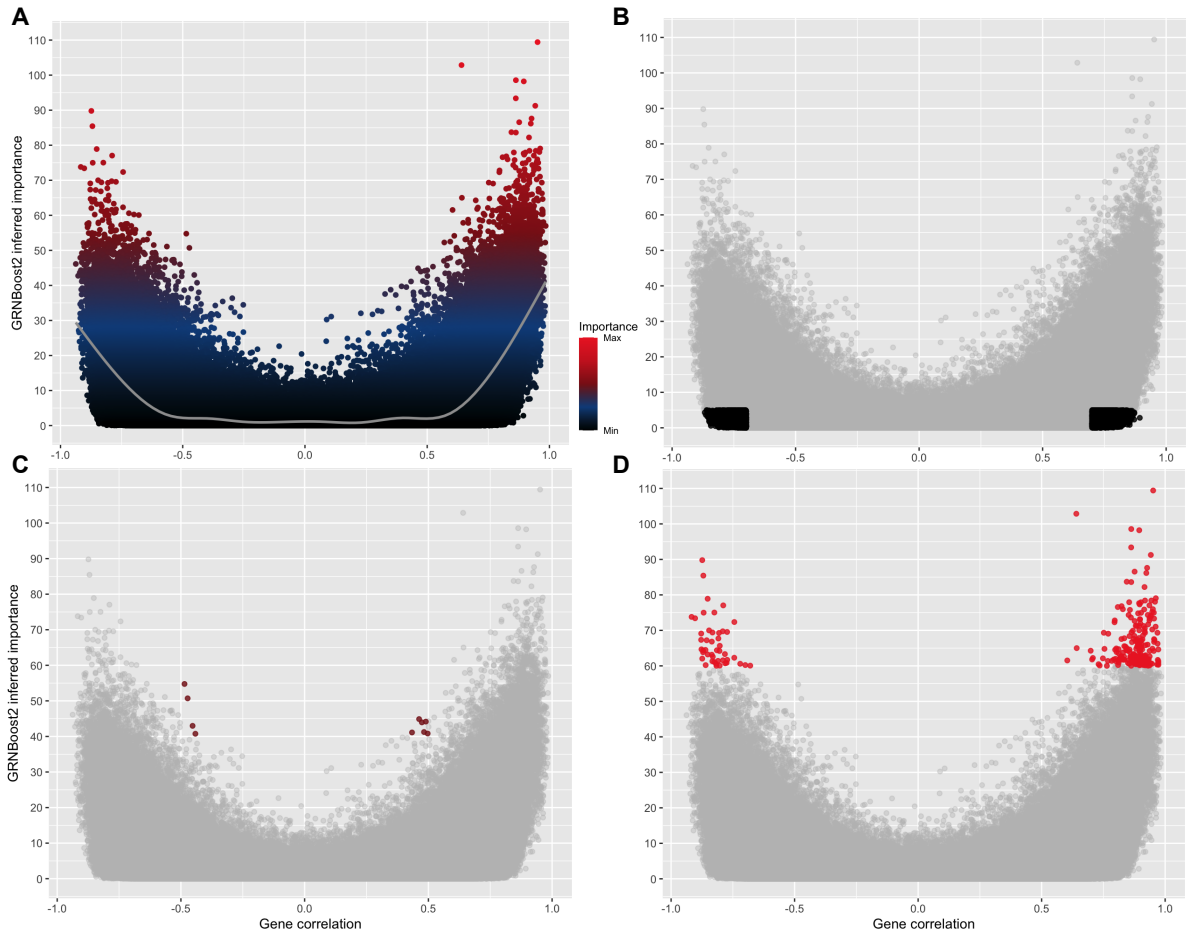


Figure 4. 2: Distribution of inferred regulatory interactions from GRNBoost2.
A) Importance values for candidate genes with their respective target genes shown in colour and y-axis. Pearson correlations for interactions on the x-axis, correlated and anti-correlated distributions shown. A grey trend line as per GAM fit illustrating a general trend between correlation and inferred importance r^2= 0.495. B) Interaction with correlations >=|0.7| with importance values <= 5 shown in black. C) Interactions with values >= 40 and correlations <=|0.5| shown in dark red. D) Interactions >= 60 and correlations >=|0.5| in red.

4.3.2 The distribution of interaction strength differs greatly between candidates

Given the sheer volume of the GRN, at over ~600 000 interactions, prioritised filtering of the network was required for further analysis. Each candidate gene (n=124) was evaluated against 5163 target genes in a one-to-one ratio, generating output values for every candidate gene with every target gene. Therefore, evaluation of candidate genes was done for only the 100 strongest responses in the network as suggested in the original GRNBoost study[79]. Ranking of interactions was done in accordance with median importance per candidate set and arranged from strongest to weakest to provide a distribution order (Figure 4.3).

Candidate gene *pf3d7_1332100* (unknown function) ranks lowest in the GRN with a mean importance of 4.31, presenting with a narrow distribution in the lowest importance ranges and therefore serving as an example of low predictive power for some candidates in the network. Candidates in the red hue regions (Figure 4.3) though numerous, serve as potentially interesting effectors of regulatory roles in *P. falciparum*. These candidates have generally larger median importance values in their top 100 interactions than the blue hue regions, which suggest these candidates possess greater predictive power with regards to their target genes. For this reason, candidates with a high importance median and known biological ability for transcriptional regulation (such as known transcription factors), will receive closer focus.

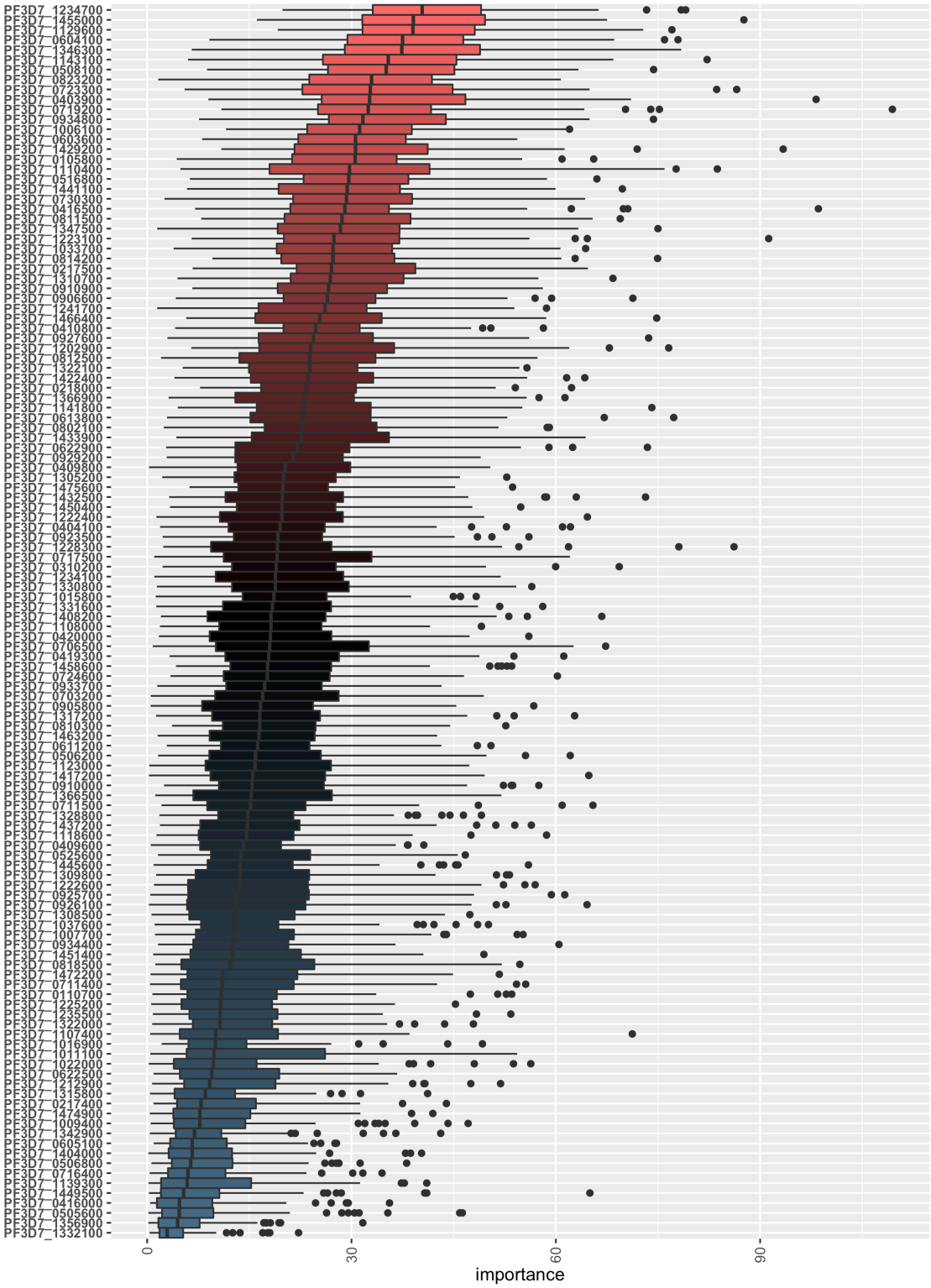


Figure 4.3: Ranked distribution of top interactions per candidate gene, prioritising candidates. Each of the 124 gene candidates have been filtered for the top 100 strongest interactions and ranked according to the sample median and represented with box-whisker plots.

4.3.3 A Focused investigation of select candidates illustrates the ability of GRNs to postulate causal relationships

After prioritised filtering of candidate genes and their subsequent quantification of the median importance metric, four candidate genes were selected for further analyses. These candidates showed strong importance values for their respective target genes and are also known transcription factors with the exception of one candidate previously identified to possess a AT-rich interaction domain (ARID) and has been implicated in gene regulation via precious studies[95]. The subset consists of *pf3d7_0516800* (*ap2-o2*), *pf3d7_0603600* (ARID domain), *pf3d7_0604100* (*sip2 -ApiAP2*) and *pf3d7_1429200* (*ap2-o3*). All but SIP2 are known for their expression during the sexual development phase of the parasite, while SIP2 expression generally peaks during trophozoite and schizont stages[48].

Expression trends for each candidate was modelled using the integrated RNA-seq dataset (Chapter 3) and displayed as smoothed lines (loess polynomial fit) accompanied by the 95% confidence intervals (CI) (Figure 4.4). Distributions of target genes, which either correlate (in red) with candidate genes or ant-correlate (blue), show clear concordance with candidate gene co-expression. It should be noted that target genes which anti-correlate, have been inferred in a predictive relationship with their respective candidate genes. It would not be possible to discuss a direct role of a repressive relationship by any of the example candidates as empirical data associated with these genes are lacking. It is furthermore clear that interactions from the network are predominantly correlated as seen by the number of positive counts (Figure 4.4).

The transcription factor *ap2-o2* is a known to increase throughout gametocyte development with peak expression in late gametocyte stages[95,105], a trend echoed in this dataset (Figure 4.4). The top target gene for *ap2-o2* (*pf3d7_1325900*, importance = 66.03) is a conserved unknown protein with no assigned function in the annotated genome of *P. falciparum*. Gene ontology for *pf3d7_1325900* would suggest a role in actin and calmodulin binding as part of its role in the microtubule associate complex, however, little is known about this protein. Interestingly, it was used in an analysis as a saliva protein marker in children between the ages of 5 and 12 for detection in subclinical parasite infections[176]. Both *ap2-o2* and -o3 where enriched for cytoskeleton remodelling upon GO analysis. The top target for *ap2-o3* is a putative alpha/beta hydrolase *pf3d7_0728700* with an importance of 93.38. A centrosomal protein CEP120 (*pf3d7_0504700*) was the second strongest target gene of *ap2-o*3 (importance = 71.95). The role of *ap2-o3* in regulating *cep120* is unclear at this point but may

serve relevance in cytoskeletal processes. A putative RNA-binding protein (*pf3d7_1126800*) with importance = 42.23, could be under the control of *ap2-o3,* however, ARID protein showed a higher importance at 51.53 for this protein and may be a more relevant regulator.

Additionally, a mRNA binding protein PUF1 (*pf3d7_0518700*) are also amongst the strongest targets for ARID with an importance value of 49.05. PUF1 is well characterised for its role in the maintenance and development of mature gametocytes. Disruption of PUF1 presents with a clear decrease in gametocytaemia after stage III gametocytes[177]. ARID may therefore be a vital component of gametocyte maturation, ARID also exhibited with strong importance values for *ap2-o3* (48.40). A concerted effort from both ARID and *ap2-o3* may be required for gametocyte maturation, particularly for their shared targets. A particularly interesting target gene of both regulators is gametocyte enriched phosphoprotein (EGXP*, pf3d7_1466200*, ARID importance = 33.04 and *ap2-o3* importance = 31.61). EGXP has previously been identified as an important antigen during gametocyte stages[178]. Individuals who with antibodies against EGXP had a lower gametocyte density (31%) over an 18 week follow up period. The disruption of these two regulators (ARID and *ap2-o3*) are possible interesting targets for transmission blocking strategies.

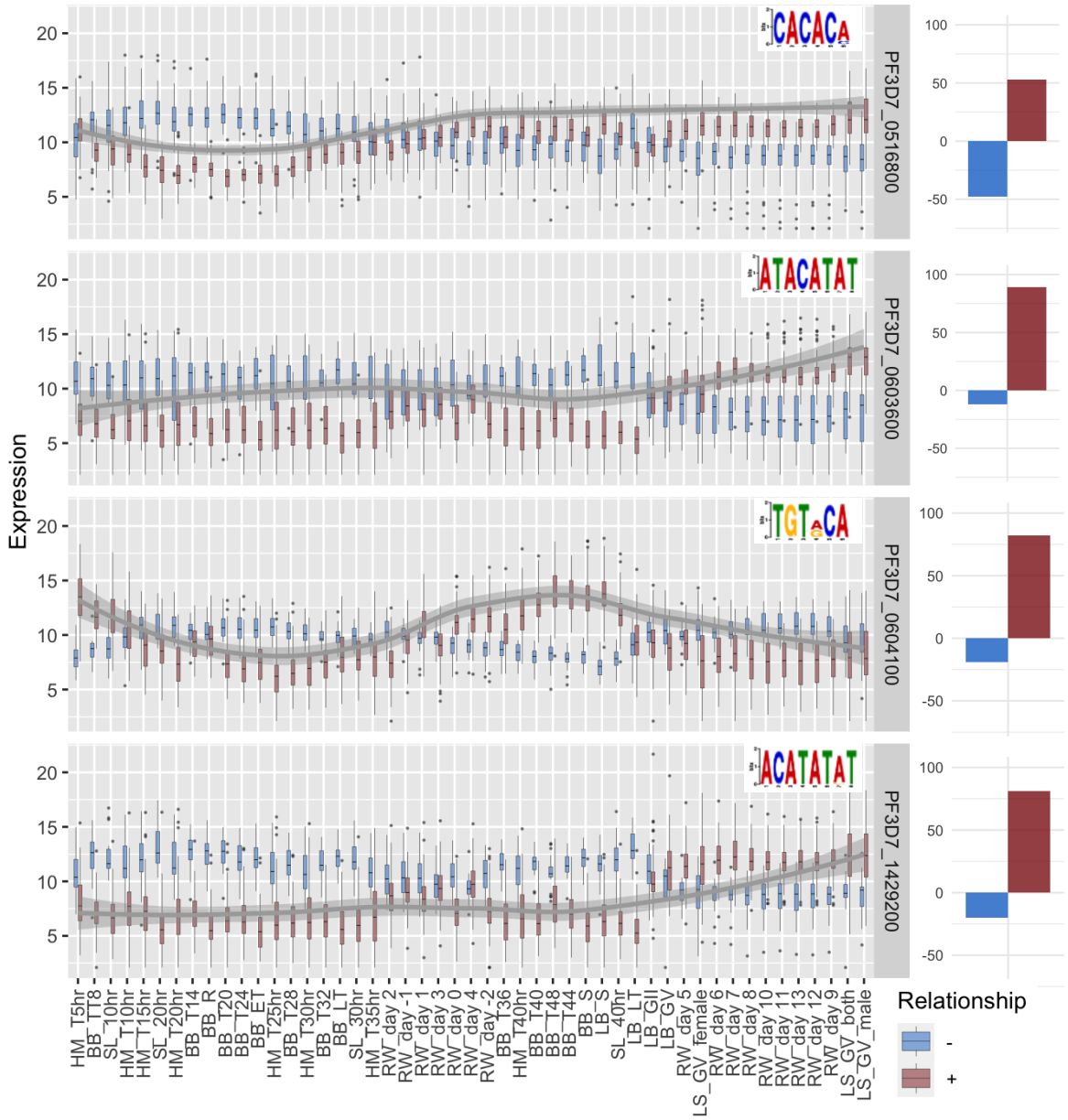


Figure 4.4: Gene Regulatory Network analysis captures strong association with candidate genes and their respective targets.
Expression as variance stabilised transformations (VSTs) are represented with box-whisker plots, separated with correlated (red) and anti-correlated (red) target subsets. Candidate gene expression is represented as line graph (95% CI grey shading) with number of correlated targets and anti-correlated indicated as bar graphs. pf3d7_0516800 (AP2-O2), pf3d7_0603600 (ARID domain), pf3d7_0604100 (SIP2) and pf3d7_1429200 (AP2-O3). DREME motifs are displayed for each candidate gene as determined by the top 100 interactions. Sample keys: BB = Broadbent, HM = Hoeijmakers, LB = López-Barragán, LS = Lasonder, SL = Siegel, RVW = van Wyk.

SIP2, a member of the ApiAP2 family of transcription factor family, has a well-documented role of transcriptional regulation during asexual development[46]. Gene ontology terms for target genes in SIP2 subset contain cytoadherence, rhoptry, apical complex, vesicle, and extracellular related processes. Not surprisingly, merozoite surface protein 10 (MSP10) is the top target gene for SIP2 in the network with an importance of 77.94. MSP10 was deemed non-

mutable in is coding domain sequence (CDS) and is considered an essential gene[179]. The transcript for a conserved unknown protein, *pf3d7_1310200*, was the second strongest target for SIP2 (75.96) and also shown to be an essential gene[179], followed by the transcript for another gene essential to asexual development, rhoptry neck 2 protein, *pf3d7_1452000*. An AP2 (*pf3d7_0613800*) is amongst the top targets of SIP2 and known to be essential. In fact, 62% of the top target genes for SIP2 are considered genes essential for asexual development in *P. falciparum*. SIP2 also shows potential regulation of *cdpk1*, which we have shown to play an important role in regulating the cell cycle in downstream signalling during the IDC[94]. Previously we determined that *pf3d7_1112100* (conserved unknown) and cAMP-dependent protein kinase regulatory subunit (*pkar, pf3d7_1223100*) may be involved in regulating CDPK1[94]. SIP2 may therefore, in conjunction with PKAr and *pf3d7_1112100* be involved in regulating CDPK1 during cell cycle progression.

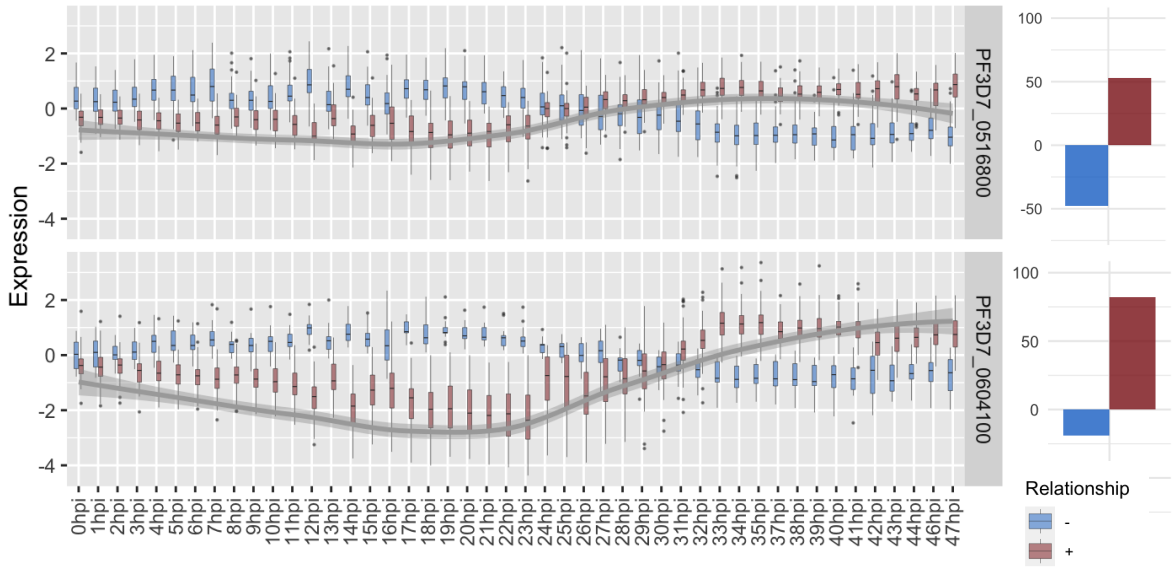


Figure 4.5: Expression profiling across the intra-erythrocytic development cycle reflect findings from GRN.
Expression is quantified as a log2(cy5/cy3) expression across a 48hr time course [48]. pf3d7_0516800 (AP2-O2) and pf3d7_0604100 (SIP2) are displayed on the Figure with correlated target genes in red and anti-correlated targets in blue. Gene counts for each candidate is shown by the bar graphs in the right-hand side of the figure. The numeric values on the x axis depicts the time at hours post invasion (hpi) as sampled in the experiment.

The GRN therefore produce a wealth of insights into the parasite biology with numerous candidates to investigate for further significance. The amount of data generated from this analysis may hold the key to future studies on candidate genes or even their respective targets. A small subset of candidates was highlighted here to prove the significance and power of this network analysis, but the overall significance can be explored much more in-depth by follow-up studies. The application of GRN analysis remains a niched subject in malaria research, and a lack of tools may be the cause for this. Although the Arboreto suite is a

complete and comprehensive python package, which proves “user-friendly” to most familiar with basic programming, this type of analysis remains out of reach for non-bioinformatics or non-computational biology researchers who are not familiar with programming. To this end this analysis framework was packaged in a “user-friendly” web-based application for researchers to easily access the tools in addition to downloading the pre-compiled networks generated form this study.

4.3.4 MALBoost: a web-based application for Gene Regulatory Network Analysis in *Plasmodium falciparum*

4.3.4.1 Intended use

GRNs constructed through supervised machine learning algorithms such as GENIE3 and GRNBoost2 offer fast and reliable tools for GRN research[71,78]. However, these tools often require familiarity with python programming and additional resources on which to run the analysis. MALBoost offers malaria researchers easy access to these machine learning tools through a user-friendly, web-based application framework. Researchers can submit their own transcriptomic data for *de novo* network construction or download a pre-compiled and validated network built for *P. falciparum* that can be used as reference framework. Users can submit their transcriptomic data along with a list of regulatory gene names to the web-application. Two options of analysis are available, GRNBoost2 and GENIE3. A “flat-file” output is generated for all the relationships along with Pearson correlations (Figure 4.6). We recommend the use of Cytoscape for easy interpretation of networks and construction of network figures. Cytoscape also hosts other secondary network analyses[180].

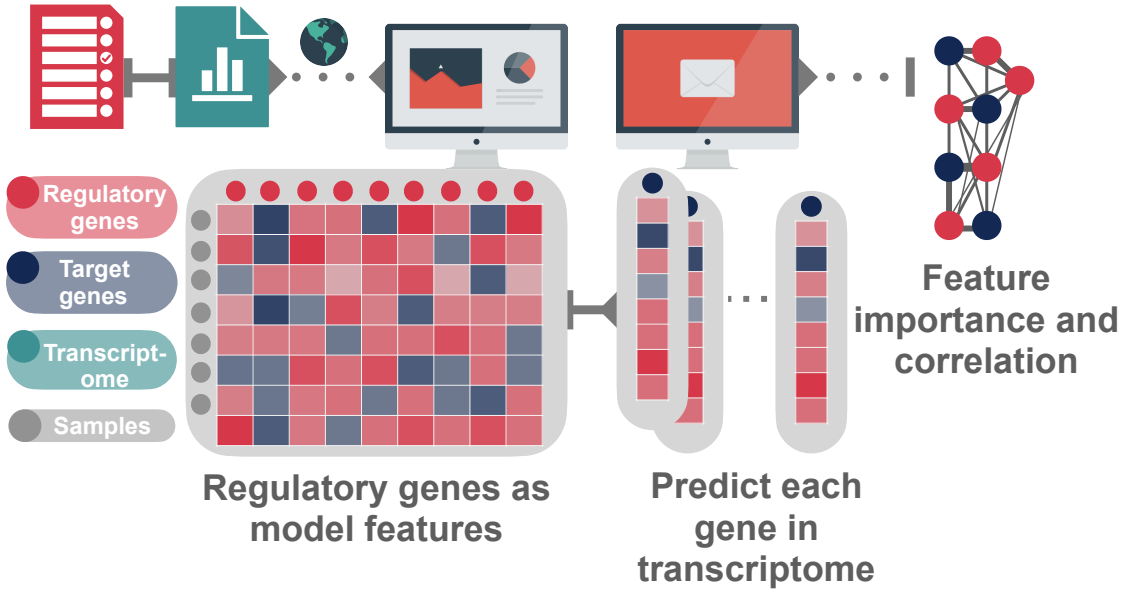


Figure 4.6: MALBoost user overview.
Users can submit their transcriptomes along with a list of regulatory gene names to http://malboost.bi.up.ac.za and select from the two algorithms available for analysis.

4.3.4.2 MALBoost architecture and interface

The application is built in a python-based Flask microframework, giving easy access to the Arboreto suite of tools [78] which runs on a CentOS VM. The application makes use of Redis data broker technology in conjunction with Celery task queuing servers to run the models in the background. Redis is an open source in-memory data structure store which allows for the funnelling of request to the queuing server Celery. The use of task queuing servers such as Celery affords the application the ability to receive multiple requests and process them on a first come first serve basis.

Request data is temporarily stored using an SQLite database. Transcriptome and regulatory list data are removed from the SQL database (DB) post submission to GRN modelling so as not to cause data congestion in the DB. Result data and user request information is stored in the DB for a period of 3 days after model completion, after which an app schedular will delete the data from the DB. The models involved in GRN construction are all implemented form Arboreto, which mask use of Scikit-Learn and Dask. A combination of HTML, CSS and JavaScript is used to render the functionality and layout of the application for the user front-end. The architecture of the application is shown in Figure 4.7, illustrating the various technologies involved.

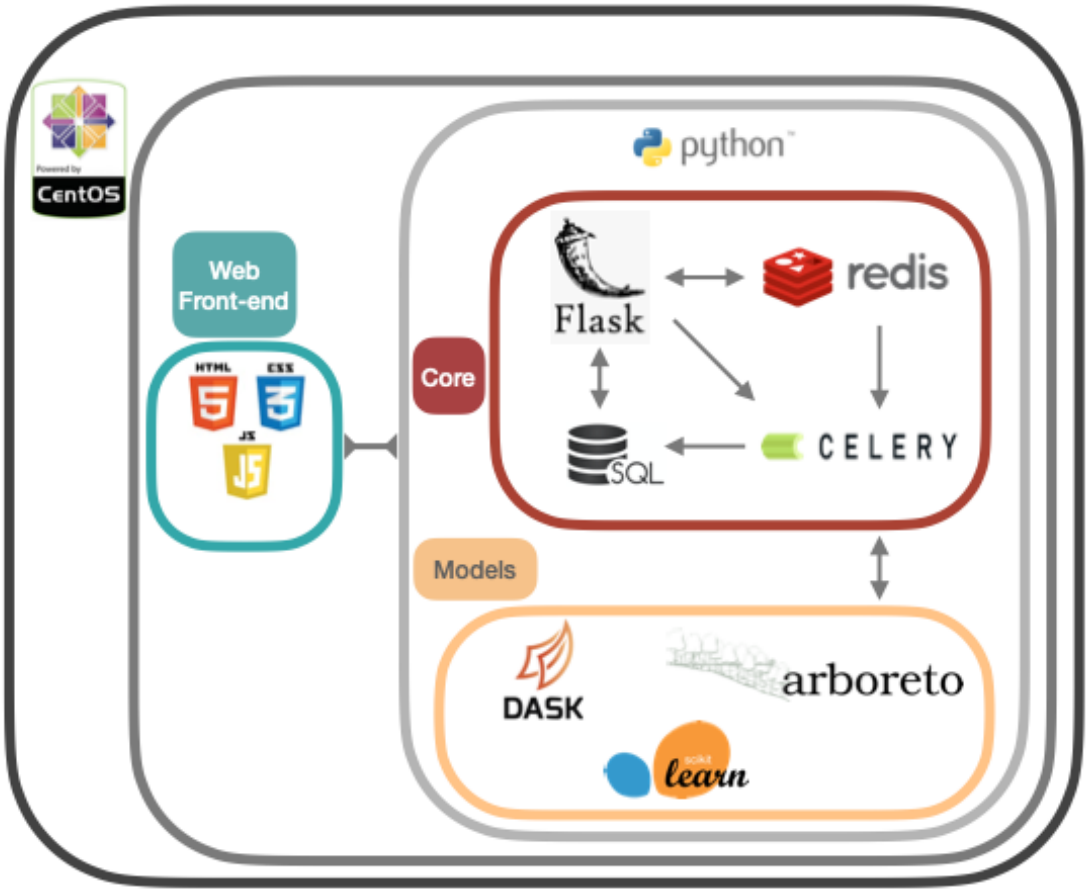
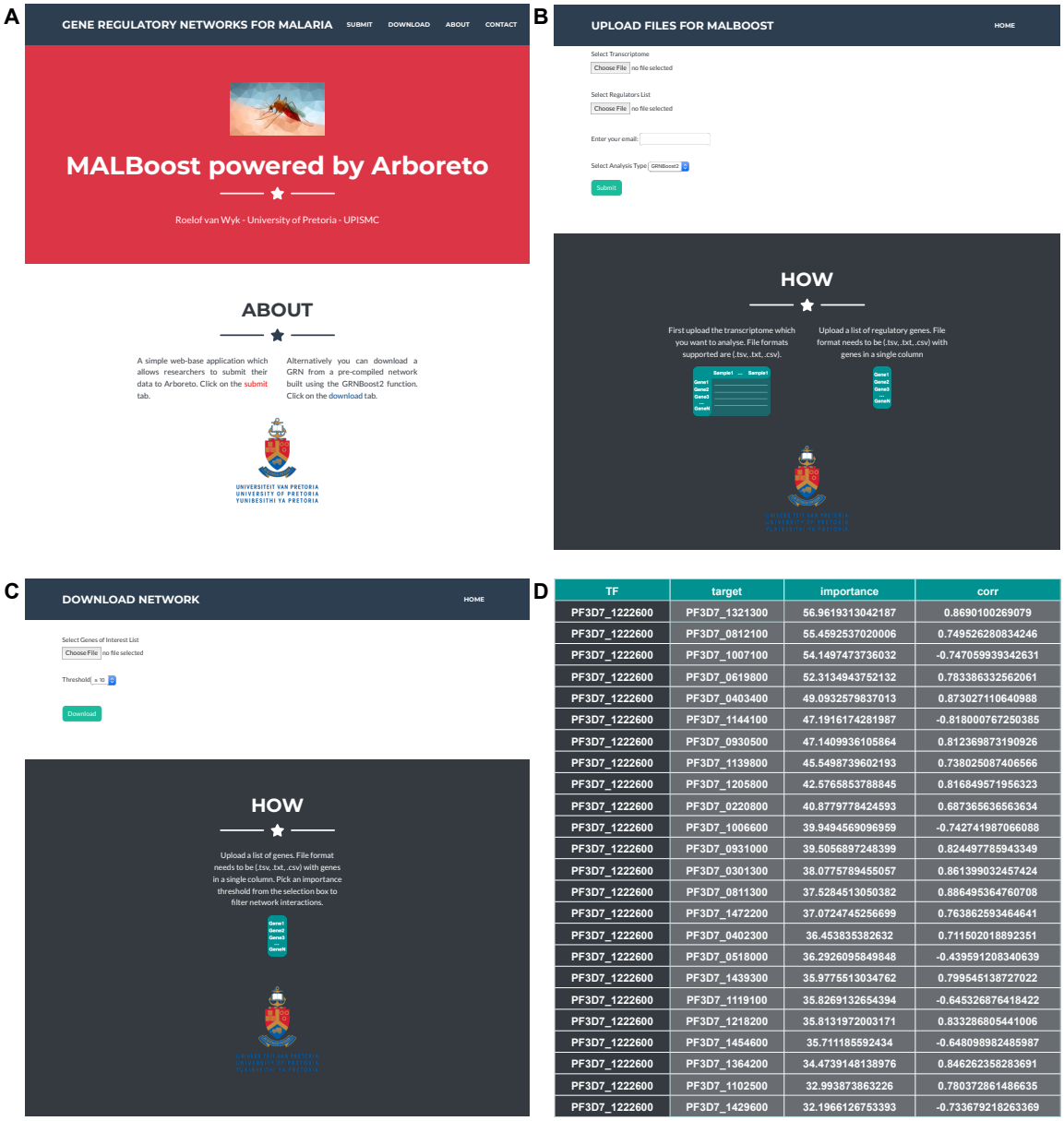


Figure 4.7: Internal architecture of MALBoost web-based application.
The application runs on a CentOS virtual machine (VM). Python formulates the core coding language of the application, running everything from Flask to task queue servers and GRN model implements. The Redis data broker passes data to the Celery queuing server, which tasks individual workers with executing the model construction. A selection of either GENIE3 or GRNBoost2 is offered, for more on the models refer to [78]. Transcriptome and regulatory list data are passed to the Celery worker environment via the SQLite DB. This data is subsequently deleted upon completion of the GRN construction, the results from the GRN is stored in the DB for a period of 3 days. Once model construction is completed a download link is provided to the researcher. The web front-end is rendered via HTML, CSS, and JavaScript.

Interfacing with the application occurs via a web portal (http://malboost.bi.up.ac.za); most modern browsers who support HTML5 should function well with the application and Google Chrome and Safari has been validated for this application. The home page of the application contains general information regarding the application as well as a navigation bar for quick navigation to functions (Figure 4.8A). Contact information and reference material for the algorithms used in GRN construction may be found on the home page. Under the submit tab, the user will be re-directed to the GRN construction page with a choice of GRNBoost2 or GENIE3 (Figure 4.8B). Two files are required from the user: a transcriptome file and a regulatory list file. The transcriptome file contains expression values from the user transcriptome with *Plasmodium* gene IDs as the index and samples as the column headers. Decision tree-based algorithms such as GBMs are often not sensitive to normalisation,

however normalisation is recommended and should be performed by the user prior to submission of their data[78]. The regulatory list file contains a list of candidate genes (gene IDs) the researcher has deemed relevant in a regulatory role and will be evaluated as such. This list has one gene ID per line of the file and most importantly these genes must be present in the transcriptome file as their values are required. Both files must be in csv, txt or tsv format. Under the download tab, the user can download a pre-complied GRN constructed with GRNBoost2 for *P. falciparum* (Figure 4.8C). A list of gene IDs of interest to the research must be supplied in previously described formats. These genes may be either candidate or target genes, the network will filter on both categories. An importance threshold is also required, which the user can assess through a drop-down tab. This will filter the network for interactions greater and equal to the set value. An example of the output data for both submit and download tabs is shown in Figure 4.8D and includes the target genes obtained associated with importance values of the interactions and Pearson correlations for each interaction in the network. The web page also includes a 'how' sections where instructions for are captured.



TF	target	importance	corr
PF3D7_1222600	PF3D7_1321300	56.9619313042187	0.8690100269079
PF3D7_1222600	PF3D7_0812100	55.4592537020006	0.749526280834246
PF3D7_1222600	PF3D7_1007100	54.1497473736032	-0.747059939342631
PF3D7_1222600	PF3D7_0619800	52.3134943752132	0.783386332562061
PF3D7_1222600	PF3D7_0403400	49.0932579837013	0.873027110640988
PF3D7_1222600	PF3D7_1144100	47.1916174281987	-0.818000767250385
PF3D7_1222600	PF3D7_0930500	47.1409936105864	0.812369873190926
PF3D7_1222600	PF3D7_1139800	45.5498739602193	0.738025087406566
PF3D7_1222600	PF3D7_1205800	42.5765853788845	0.816849571956323
PF3D7_1222600	PF3D7_0220800	40.8779778424593	0.687365636563634
PF3D7_1222600	PF3D7_1006600	39.9494569096959	-0.742741987066088
PF3D7_1222600	PF3D7_0931000	39.5056897248399	0.824497785943349
PF3D7_1222600	PF3D7_0301300	38.0775789455057	0.861399032457424
PF3D7_1222600	PF3D7_0811300	37.5284513050382	0.886495364760708
PF3D7_1222600	PF3D7_1472200	37.0724745256699	0.763862593464641
PF3D7_1222600	PF3D7_0402300	36.453835382632	0.711502018892351
PF3D7_1222600	PF3D7_0518000	36.2926095849848	-0.439591208340639
PF3D7_1222600	PF3D7_1439300	35.9775513034762	0.799545138727022
PF3D7_1222600	PF3D7_1119100	35.8269132654394	-0.645326876418422
PF3D7_1222600	PF3D7_1218200	35.8131972003171	0.833286805441006
PF3D7_1222600	PF3D7_1454600	35.711185592434	-0.648098982485987
PF3D7_1222600	PF3D7_1364200	34.4739148138976	0.846262358283691
PF3D7_1222600	PF3D7_1102500	32.993873863226	0.780372861486635
PF3D7_1222600	PF3D7_1429600	32.1966126753393	-0.733679218263369

Figure 4.8: MALBoost web-based application for GRN construction in *P. falciparum*.
A) Website home page with navigation bar and description of the application. Contact detail, about and a how to guide is found on the home page along with reference material to the GRNBoost2 algorithms paper http://malboost.bi.up.ac.za. B) The submit tab where researchers can construct their own GRNs base either on GRNBoost2 or GENIE3. Submission of transcriptome as well as a list of candidate genes (suspected in a regulatory role) is supplied as either tsv, txt or csv format files. C) Download tab where researchers can download results based on a pre-compiled GRN. Researchers supply a list of gene IDs which they are interested in and apply an importance threshold which will download the resulting network. D) Results format from the network (either constructed or pre-compiled). TF = Transcription Factor or candidate gene, target = target gene, importance = importance value assigned by the selected model, corr = Pearson correlation of the interaction.

4.3.4.3 MALBoost usage in GRN construction for transcriptional regulator

To validate the accuracy of the tool, MALBoost was interrogated with a known gene regulator, the AP2-G transcription factor, which was used as candidate and for which the targets were

then probed in the tool. AP2-G is a known regulator of sexual commitment in *Plasmodium* parasites[37,181] and was found to bind upstream of a specific subset of *P. falciparum* genes and associated with an increase in the transcript abundance of these genes[38]. Here, a GRN was constructed from 49 RNA-seq samples encompassing both asexual and sexual blood-stage development (Chapter 3) and the target genes of AP2-G were extracted from the GRN output (Figure 4.9). Using >30 importance as a highly significant guided threshold resulted in the identification of 28 putative targets, of which 5 have empirical ChIP-seq data showing their promoter regions are bound by AP2-G. This constitutes a significant overrepresentation (P<0.05) of known AP2-G targets bound in sexually committed ring-stage parasites and two of these genes were also differentially expressed following AP2-G genetic perturbation (Figure 4.9A&B). Comparatively, using Pearson correlations to investigate showed the 28 most correlated transcripts with AP2-G over the full transcriptome only contained 3 ChIP targets and none were perturbed following AP2-G knock down. The top 100 interactions included a further 11 AP2-G ChIP targets, but these were not significantly enriched above a random chance of occurrence (Figure 4.9B). Interestingly, targets that were bound uniquely by AP2-G rather than in overlapping regions with a second transcription factor, AP2-I, were also overrepresented although not quite significant (P<0.1). However, targets of AP2-I either shared with AP2-G or unique to AP2-I were not overrepresented in the putative targets (Figure 4.9A). These results suggest that MALBoost could pick out some directly regulated transcripts within bound target genes of transcription factors.

While the AP2-G bound genes are of interest for direct genetic regulation by the AP2-G transcription factor, probing into the top target genes (>30 importance) of AP2-G also yield interesting implication for downstream regulation. In addition of the 5 directly-bound AP2-G ChIP targets, a further 7 of the 28 putative targets were transcriptionally effected following a genetic knock-out of EBA175 (*pf3d7_0731500)* (Figure 4.9C), an AP2-G ChIP target which regulates parasite invasion[182]. Furthermore, the putative targets were primarily (16/28) transcriptionally abundant in gametocytes (Figure 4.9C) the stage in which AP2-G is expected to regulate gene expression and it is possible that a factor downstream of AP2-G also regulates these targets directly, i.e. an epigenetic regulator (*hda1, pf3d7_1472200*) or post-transcriptional regulator (*caf1, pf3d7_0811300*). This suggests that MALBoost was able to find many real targets of sexually committed parasites even in bulk RNA datasets.

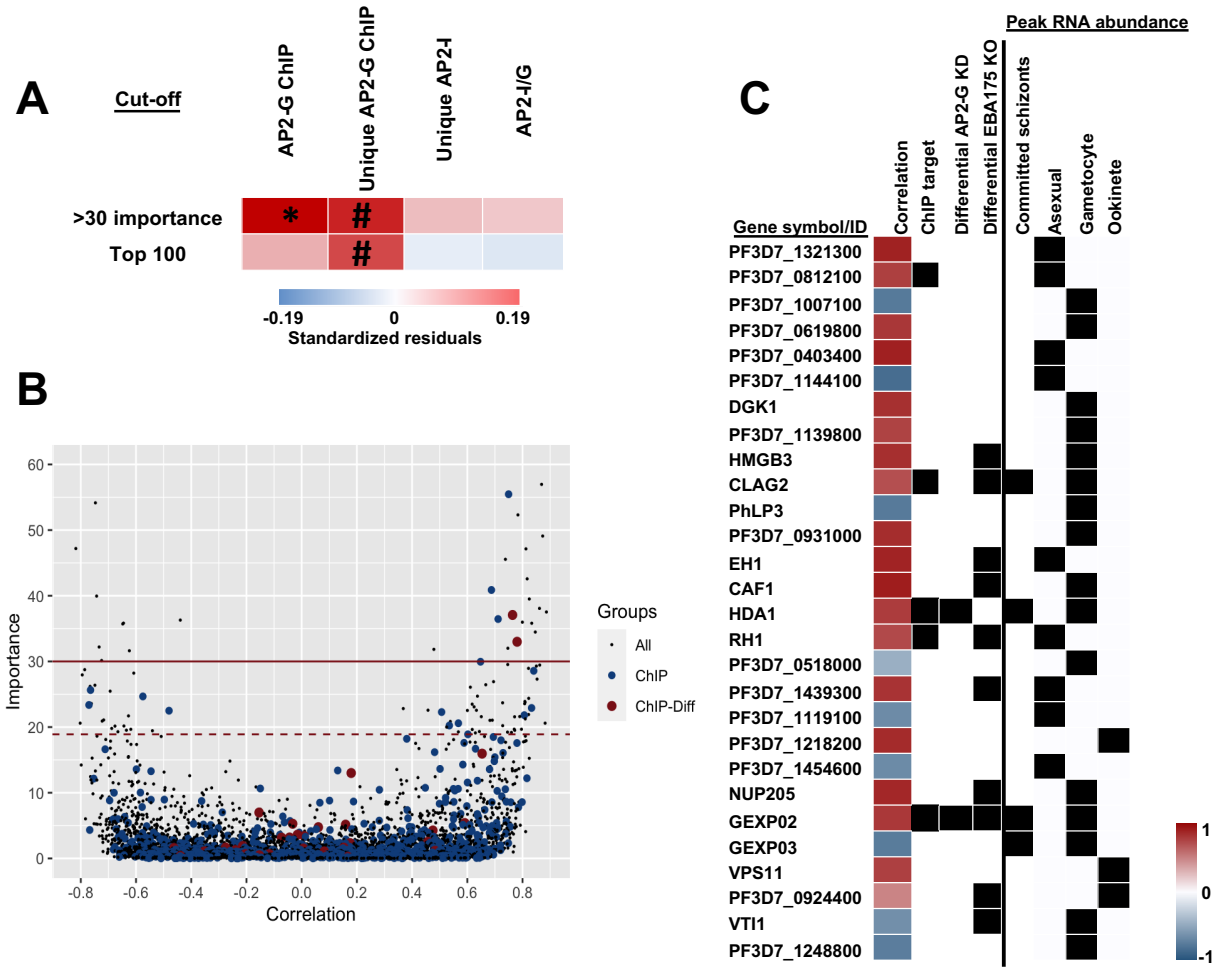


Figure 4.9: MALBoost results for AP2-G GRN.
A) The number of ChIP-targets for AP2-G and AP2-I that were within the top putative target genes of AP2-G either by importance or using top interactors were tested for significant overrepresentation using a two-tailed Fisher's exact test (#=*P*<0.1, *=*P*<0.05). B) Distribution of importance over Pearson correlation for the investigated genes (5142) with ChIP-targets (ChIP) and ChIP-targets that were differentially transcribed following AP2-G knock down (ChIP-diff) highlighted. Solid line=>30 importance threshold, dashed line = Top 100 interactors. C) Genes above the 30 importance threshold were ordered by decreasing importance and Pearson correlation shown along with hits in AP2-G ChIP-seq, differential transcripts in the AP2-G knockdown, EBA175 KO, and the stage at which the transcripts peak in abundance.

4.4 Discussion

The use of GRN analysis in *Plasmodium* research has not been frequent while most fields have benefitted from this line of inferential reasoning. Many experimental investigations regarding regulatory mechanism and co-expression studies have been conducted, but the utilisation of GRNs remain scarce. Previously we used DBNs and weighted correlation networks to gain some understanding and insight into transcriptional regulation, particularly during gametocyte development, but not exclusively (Chapters 2&3). Many of these findings resolved some specific regulatory mechanisms such as the role of Ca^{2+} signalling cascades

with assorted kinase interconnected with transcription factors such as MYB1, ApiAP2's as well as epigenetic modifiers such as SET9 and SAP18 for their roles in cell cycle proliferation (Chapter 2). We established key transcription factors and target genes during gametocyte development as well, with some level of inference regarding uncharacterised DNA-binding proteins, which may be involved in regulation (Chapter 2&3). Two notable drawbacks were encountered during these analyses which are unique to either DBNs or Co-expression networks. Co-expression networks seldomly deal with predictability of the relationship[69,164,183,184]. The networks capture a relationship but cannot accurately define a plausible causal root. It is merely noting the co-occurrences of transcripts which may have significant relation to one another or not. DBNs on the other hand do establish an element of predictability which would ascribe some root causal inferences[73,76,92], however these networks are computationally exhaustive to perform at scale. This resulted in the testing of a small pool of target genes and an even smaller pool of regulatory candidates. Evaluating the full transcriptome with the maximum suspected pool of regulatory candidates, was computationally sub-optimal.

Increasing our scope for transcript regulation inference with the inclusion of the largest group of candidate regulatory genes and the entire transcriptome, required the use of supervised machine learning techniques such as GRNBoost2. The advances of GRNBoost2 in terms of speed and size of analysis greatly exceeded that of DBNs, with the added function of predictability inference which is lacking in co-expression networks[78]. The construction of a global GRN as a result of a single analysis capturing both developmental phases would within reason produce directly comparable results for all the interactions generated.

Heterogenous sampling is often used for these kinds of networks and in particular for GRNBoost2 (then version 1)[79] matching the sampled datasets used here. Sample size remains a concern when dealing with machine learning algorithms. Gradient boosting algorithms have been shown to be successful with samples sizes as small as 25[185], though the added effect of sample heterogeneity and size is sure to reduce the power of the evaluation. The conventional wisdom posits that the greater the sample size the more reliable the results. Unfortunately, not many RNA-seq gametocyte datasets exist for *P. falciparum*, which is why our in-house data (Chapter 3) was generated to address some of the shortage. Fortunately, DNA-microarray datasets always serve as an independent sanity check regarding patterns and targets derived such as with Figure 4.5.

A minor concern for this study was the performance of GRNBoost2 on the relevant dataset with regards to the importance scores and the assumption of predictability. Since the metric

produced from the analysis is essentially a modified feature importance score, we do not have a clear sense of the model performance in each iteration for predicting the outcome. Meaning if we train a model of 124 genes on target x, how well did the model manage to predict the outcome? This validation was done through re-constructing the same Scikit-learn based GBM models using the same parameters as for GRNBoost2 and calculating the K-fold cross validation sets. A normalised root mean square error (NRMSE) was used alongside a 95% prediction interval (data not shown). The inferred importance value appears to have captured this error as the error reduced with increase in importance. This validation step re-capitulates the original assumptions regarding GRNBoost2 that high importance scores directly relate to lower error rates, thus are relevant to the accurate prediction of target outputs.

The global GRN quantified a total of 636 732 interactions for the 124 candidate regulatory genes. This vastly exceeds what we were able to quantify with DBNs during the cell cycle and gametocyte studies in Chapter 2. This quantification was also achieved off server on a desktop computer (Mac mini, 3 GHz 6-Core Intel Core i5, 32 GB DDR4 RAM), with runtime in 9.24 minutes. This is compared to 48 h runs performed on the DBN studies. This shows the GBM based approach is more feasible for GRN construction as this drastically advances the data processing and allots more time for analysis. The GRN also showed gains over co-expression networks, with co-expression often failing the element of predictability as observed in Figure 4.2. The analysis shows correlation “loosely” associating with higher predictability (importance score), however numerous relationships which correlate strongly were unable to present predictability Figure 4.2B. This is an expected outcome as correlation does not equal causation, therefore the GRN excavates several more promising interactions.

AP2-O2 was a strong feature for the network, impacting primarily on genes involved in cytoskeletal/actin process, suggesting that this TF may be important for the development and progression of gametocytes. Given the expression of this TF well into stage V gametocytes, it may well be required to continue the cytoskeletal changes required to progress through these morphologically distinct stages[166]. AP2-O3 also showed involvement with genes in the cytoskeletal process. AP2-O3 and ARID transcripts appear to share in a consorted effort with the regulation of PUF1, *pf3d7_1126800* (RNA-binding protein) and EGXP. The inhibition of PUF1 also critically affects progression from stage III onwards[177]. EGXP is also relevant for its role in host-parasite immunity. Patients with anti-EGXP antibodies showed a 32% reduction in gametocyte volume[178]. It is possible that ARID and AP2-O3 are both required in their regulatory roles for a subset of genes, were AP2-O3 would act as the primary TF and ARID assisting in the role of transcription.

Utilising the power of the combined dataset produced in Chapter 3 as well as funnelling the relevant regulatory candidates across three studies (Chapter 2&3), we could introduce an added level of complexity in the form of predictive modelling. The power of the predictive modelling expands our knowledge beyond the "guilt-by-association" trappings found in correlation-based metrics and resulted in greater scalability than the DBN research previously conducted (Chapter 2). This comprehensive GRN denotes the largest of its kind in the field to date and the potential for numerous investigations beyond the examples highlighted here are interesting research prospects.

Access to tools for non-Bioinformatic based researchers often proves the "hurdle". Here, we endeavour to provide greater access to tools such as those in the Arboreto suite, with the potential for tool upgrades in the platform. This provision is setup through a web application interface named MALBoost after the GRNBoost-based architecture at the core of the application. The tool is intended for use by Malaria transcriptomics researchers and the provision of a pre-compiled GRN, the very network constructed in this chapter.

The application framework focused on ease of implementation and maintenance. I built a Python-Flask application to provide both advantages as Python-Flask is a popular and well documented framework. This would allow future upgradable components to be written as simple python scripts which essentially "plug-in" to the application, making the addition of new tools simple. Django may also be considered a suitable alternative to Flask, however, requires a more extensive setup than Flask. Flask applications are widely used in industry "Apps" catering to the needs of mobile "back-end" applications and often used as a framework for interfacing with databases [186]. Utilisation of a python-based Flask microframework ensures easy deployment and the combination of Redis and Celery technologies provide a powerful backend capable of running the computationally intensive Arboreto suite. The core software of the application (Arboreto and the recommended use of GRNBoost2) has shown tremendous promise within the field of cancer research [78,79] and should translate well to use with the comparatively small genome and collective datasets in the malaria research field. Here we produce and apply a web-based application of this technology to extant malaria transcriptomic and ChIP-seq datasets.

The ApiAP2 sequence-specific family of DNA binding proteins provide some of the only probable evidence of specific transcriptional regulation in *P. falciparum* parasites. The AP2-G transcription factor is one of the most critical proteins for study as it is essential for progression into sexual differentiation. Previous data also show that this protein directly binds the nuclear genome at very specific sites[37,38] and influences the transcript abundance of specific genes.

Within the very narrow subset of genes that are both directly bound and putatively regulated by this transcription factor, our application showed 5/28 genes could be predicted by transcriptional profile in bulk RNA-seq datasets alone. Interestingly, while this represents an overrepresentation of AP2-G targets in the predicted dataset, the algorithm did not identify over-representation of AP2-I in the predicted targets of AP2-G, despite 35% of AP2-G genome regions also being bound by AP2-I[10]. In addition, probing into the unbound predicted target genes also provided interesting results as 7 more were possibly indirectly affected, as they were transcriptionally perturbed following genetic knock out of EBA175 (*pf3d7_0731500*), a gene regulated by AP2-G[182]. Overall, these results suggest MALBoost can discern direct and indirect effects of transcriptional regulators within a complex sample of transcriptome datasets and provides independent analysis of transcriptional variation that can be explained by the expression of a singular transcriptional regulator within complexly regulated genesets. The main intended use of this application would be to assist in researcher "think tanks" or "brainstorming sessions" whereby researchers prioritise genes they which to experimentally investigate through construction of custom GRNs. This effectively creates a pseudo-simulation environment for quick testing and experimental prioritisation strategies.

Chapter 5

Concluding discussion

The *P. falciparum* parasite is well known for its complex life cycle and a wide array of developmental phases[166]. The molecular drivers behind this complex development remains a relevant research focus, given that for instance, so few transcription factors have been identified for *P. falciparum* that could explain the gene expression patterns underlying the different stages of development. It would also stand to reason that the chromatin landscape is expected to impact regulation during gametocyte development, however this landscape is only partially characterised[134,187,188]. Post-transcriptional regulation mechanisms are also only partially understood, though their involvement is certainly implied as exhibited with mRNA decay studies[48,50]. The added complexity of non-canonical regulatory mechanisms such as lncRNA and how little is known for gametocyte development present with a knowledge gap in our understanding of gametocyte biology[42]. We addressed some of these knowledge gaps in various ways, primarily through GRN analysis applied to different life cycle phases.

In chapter 2 we explored the useability of DBNs for both asexual and sexual phases of development, resolving numerous regulatory elements for each phase. We identified early regulators involved in the G1/S phases transition and molecular regulators of a newly identified cell cycle checkpoint. Reconstruction of the sequential events following cell cycle progression post the G1/S phase and cell cycle re-entry, proved to be a complex series to resolve. The DBN proved an appropriate choice for such a complex problem as they are purposely built to resolve time based sequential dependency in the data. Similarly, sexual development which is also a time-based progression process, benefited from using DBN analysis and time course data. Interrogating the relevance of ApiAP2 transcription factors during gametocyte development using the DBN approach, resolved more robust associations than prior attempts to deconvolute the dataset[105].

The ability of DBNs to capture the directionality of the relationships also produce fruitful interpretations as repressive interactions are often hard to determine and ascribe, particularly the role of *ap2-g2* and *pf3d7_0611200* and repression of asexual transcripts during gametocyte development. Although DBNs are often difficult to construct or execute, their interpretation is often intuitive and easy to perform. This greatly increases the speed of conclusions during analysis and interpretations of the data become easy to contextualise. The most significant drawback with using the DBN approach, was scalability of the analysis.

The scalability and speed constraints may in part be due to the R build not being suitable for larger datasets, however, the algorithm itself proves a computationally expensive one[73,75,92]. Opting for a DBN build in a compiler language may have aided speed and scalability, but the algorithm does have some inherent scale/speed issues which are well documented[73]. DBNs exist in many different configurations, using different Bayesian structure search algorithms with GRENITS employing a MCMC approach. Improvements to the DBN approach exist in the form of reversible jump Markov Chain Monte Carlo (RJMCMC) which GRENITS does not use, however the use of Gibbs Variable Selection by the algorithm does aid in the sampling from correct conditional distribution[73,109]. Other additional components have been added to DBNs over the years such as the use of Bayesian Principle Component Analysis (BCPA), which is more relevant to imputing missing values (not relevant to this work)[73]. All things considered the use of GRENTIS proved to be fruitful in our analyses, even though the scalability suffered during the course.

Following our newfound understanding of the different life phases of the parasite development in human hosts, we aimed to contrast the developmental phases against one another in a balanced way, rather than studying the phases in isolation. Integrating data across experiments and platform differences is often difficult. To circumvent this issue, we chose appropriate datasets and tested several normalisation methods before settling on VST, that produced the most comparable distribution between samples. In addition, as the bulk of RNA-seq experimental datasets available were conducted on the asexual phase of development, we contributed a full time-course gametocyte bulk RNA-seq dataset to enrich the data for gametocyte development which we combined with relevant asexual proliferative RNA-seq data in a balanced manner, as to not overrepresent on phase over the other. We could then accurately contrast and compare the asexual and sexual phase of *P. falciparum* development to interpret elements which seem conserved between phases, unique for phases or which are in flux during various stages across both phases.

This level of resolution is often missed when studying one phase in isolation as many of the genes prove to be active in other phases of development and this often limits the conclusions of studies. Using a balanced, contrasting dataset in an unsupervised model such as the weighted co-expression network provide many interesting self-organising relationships for inquiry. Many of the genes required for RT stages for example, were often required during gametocyte stages as well, with these roles highlighted more clearly in this model than our prior work which focused on one phase at a time. It became clear that genes which are crucial in gametocytes, may also have functional roles in other stages. It is easy to overlook these

genes which appear at first glance to be redundantly expressed across multiple stages. However, several of these shared genes appear to be strong candidates in the gametocyte subnetwork, including ApiAP2 DNA binding proteins, that are often phenotypically essential at a specific point in development, despite being expressed in multiple life cycles stages[41,134]. It is then possible that this observation might hold true for other regulators identified from the subnetwork.

As a higher-level evaluation of transcriptional control, we investigated neighbouring gene pair expression patterns, with 286 genes pairing in a co-expressed manner which either fit a head-to-head or tail-to-head configuration. The hypothesis would follow that gene pairs in this configuration which strongly co-express would possess a greater likelihood of sharing a promoter. Although independent transcription always remains a possibility for these pairs, it would stand to reason that shared promotors may be more likely, particularly so for head-to-head configurations. These head-to-head configurations have generally short distance between gene start points with the existence of bi-directional promoters possibly explaining the strong correlation between these gene pairs. Experimental confirmation regarding shared promoters would prove an interesting future research prospect.

The discovery of potentially novel lncRNA in gametocytes and their related genes during mature gametocyte development posits an interesting avenue for future research. The role of lncRNA in gene regulation, has been characterised as both gene silencing or gene activating in *P. falciparum*. The most prominent example of transcriptional level gene regulation is; telomeric/subtelomeric *var* gene silencing using intron-derived lncRNA ultimately resulting the recruitment of histone-modifying enzymes such as *pfkmt1* which generates heterochromatin via H3K9me3 mark deposits[63]. Conversely activation of *var* genes occurs (internally located) through the use of lncRNA[189]. Here intron-derived antisense lncRNA was used in an exogenous plasmid to activate *var* genes. The juxtaposition between these activation and repression mechanisms, would appear to be explained by the location of the *var* genes themselves, internal vs telomeric[42]. These findings show that lncRNA and their gene regulatory effects are very complex and difficult to extrapolate without direct evaluation. This presents with difficulties in interpreting the lncRNA co-expression data, as both repression and activation remain possibilities. It's therefore not clear whether co-expression or anti-correlation patterns would be evidence of activation or repression or indirect/downstream effects.

Regulatory candidates derived as outcomes from DBN and WGCNA studies, informed our strategy to construct a more sophisticated network with GRNBoost2. This approach was

beneficial as we could scale the analysis to include all the proposed candidate genes, as well as all available target genes shared across datasets. What was clear from the preliminary results of the GRNBoost2 analysis is that the relationships went beyond “guilt-by-association” assumptions as found in correlations. The WGCNA does attempt to reduce some of these biologically coincidental relationships through accounting for the topological scale and transforming to a scale free topology. However, the calculations at the core of the approach are not directional in nature and will not indicate if the relationship is causal. Tools that attempt to predict the outcome as part of their framework, retain the ability to reconstruct the observable data to an extent and make for easier interpretations regarding causal relationships. These reconstructions at the root level posed an important question regarding the use of GRNBoost2, which was the validity of the individual model constructions themselves.

Given that GRNBoost2 uses an iterative approach to evaluate the model (regulatory genes) as x with each individual target gene (y) across all samples, the performance of each iteration needed to be evaluated rather than assumed. If the model was poor at predicting the target gene in the iteration step, then the interaction scores would have been assumed to be low. We evaluated this by reconstructing the iterative process (using similar parameters) and calculated the NRMSE and 95% prediction intervals for each target gene and concluded that the low interaction scores associated with worse predictions. Thus, holding the original assumption that high importance scores will correlate with more reliable models for GRNBoost2 performance and appropriateness.

This importance score was informative in prioritising the candidate genes in the network, not just on the basis of individual interactions, but overall interactions for each candidate as some candidates appeared to have a larger role in regulation. In addition to prioritisation of candidate genes, the use of an established tool provides some confidence in uncharted analysis where little is known about *P. falciparum* gene regulation at the interaction level. Secondary benefits of using ensemble learning algorithms is their noteworthy lack of sensitivity to normalisations[185]. In principle this would mean that the algorithm would produce the same findings pre- and post-normalisation. This was not a point of contention for our network as we used a robust normalisation strategy across multiple datasets which perform well in co-expression analysis but is indeed an added benefit of using this approach.

Ultimately with principled level understanding of this approach, researchers can make great strides in their own research. Particularly those who seek to understand the influence of

regulatory genes on other genes during any given developmental phase. The constraint often presents itself in a lack of familiarity with scripting languages, although GRNBoost2 is a very comprehensive library, a certain amount of Python knowledge is required to run the analysis. It is therefore important to make tools of this nature as broadly accessible as possible. This often requires simple graphic interfaces to give users the power of the analysis without having to learn an entirely new discipline. This is what we aimed to achieve with MALBoost. The framework offers researchers the ability to effectively simulate experiments by predicting the effect of their proposed regulatory candidates on target genes. This could prove to be a major optimising step in experimental design. Improvements regarding the MALBoost framework is also expected to be implemented over time. Expansion of the analysis repertoire such as the inclusion of DBN (for small time courses), basic network statistical breakdowns and visualisation options may greatly improve the user experience adding greater value to their research.

In conclusion, we explored several interesting approaches to construct GRNs for the sexual and asexual developmental phases of *P. falciparum*, resulting in a detailed description of regulatory events that shape the parasite's life cycle. We used DBNs to deliver reliable and robust results by asking specific, small-scale biological questions which allowed us to identify a few key regulators that can be further investigated. We then probed an unsupervised correlation-based approach that could be adopted at large scale which was very useful in making between stage and phase comparisons, but this correlation had its own limitations in that the relationships often lacked explanatory power. Ultimately, we constructed an ensemble based GRN as the most favourable for in-depth analysis at scale. This final approach allowed for a fine separation of molecular regulators involved in sexual, asexual development or both with high accuracy and at large scale. We believe that the candidates and processes we highlight in this thesis constitute numerous avenues for future research in way of the novel regulators that can be characterised in future studies.

Reference:

1. World Health Organization. *World Malaria Report*. (2020).
2. Datoo, M. S. *et al.* Efficacy of a low-dose candidate malaria vaccine, R21 in adjuvant Matrix-M, with seasonal administration to children in Burkina Faso: a randomised controlled trial. *Lancet* **397**, 1809–1818 (2021).
3. Famin, O. & Ginsburg, H. Differential effects of 4-aminoquinoline-containing antimalarial drugs on hemoglobin digestion in *Plasmodium falciparum*-infected erythrocytes. *Biochem. Pharmacol.* **63**, 393–398 (2002).
4. Asawamahasakda, W., Ittarat, I., Pu, Y. M., Ziffer, H. & Meshnick, S. R. Reaction of antimalarial endoperoxides with specific parasite proteins. *Antimicrob. Agents Chemother.* **38**, 1854–1858 (1994).
5. Tse, E. G., Korsik, M. & Todd, M. H. The past, present and future of anti-malarial medicines. *Malar. J.* **18**, 93 (2019).
6. Risco-Castillo, V. *et al.* Malaria Sporozoites Traverse Host Cells within Transient Vacuoles. *Cell Host Microbe* **18**, 593–603 (2015).
7. Arnot, D. E., Ronander, E. & Bengtsson, D. C. The progression of the intra-erythrocytic cell cycle of *Plasmodium falciparum* and the role of the centriolar plaques in asynchronous mitotic division during schizogony. *Int. J. Parasitol.* **41**, 71–80 (2011).
8. Alleva, L. M. & Kirk, K. Calcium regulation in the intraerythrocytic malaria parasite *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **117**, 121–128 (2001).
9. Garnham, P. C. C. & Kendall G., P. Periodicity of infectivity of plasmodial gametocytes: The “Hawking phenomenon”. *Int. J. Parasitol.* **4**, 103–106 (1974).
10. Josling, G. A., Williamson, K. C. & Llinás, M. Regulation of Sexual Commitment and Gametocytogenesis in Malaria Parasites. *Annu. Rev. Microbiol.* **72**, 501–519 (2018).
11. Portugaliza, H. P., Llorà-Batlle, O., Rosanas-Urgell, A. & Cortés, A. Reporter lines based on the gexp02 promoter enable early quantification of sexual conversion rates in the malaria parasite *Plasmodium falciparum*. *Sci. Rep.* **9**, 14595 (2019).
12. Silvestrini, F. *et al.* Protein Export Marks the Early Phase of Gametocytogenesis of the Human Malaria Parasite *Plasmodium falciparum*. *Mol. Cell. Proteomics* **3**, 1437–1448 (2010).
13. Pelle, K. G. *et al.* Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. 1–20 (2015) doi:10.1186/s13073-015-0133-7.
14. Macrae, J. I. *et al.* Mitochondrial metabolism of sexual and asexual blood stages of the malaria parasite *Plasmodium falciparum*. **11**, 1 (2013).
15. Hawking, F., Wilson, M. & Gammage, K. Evidence for cyclic development and short-lived maturity in the gametocytes of plasmodium falciparum. *Trans. R. Soc. Trop. Med. Hyg.* **65**, 549–559 (1971).
16. Hanssen, E. *et al.* Soft X-ray microscopy analysis of cell volume and hemoglobin content in erythrocytes infected with asexual and sexual stages of *Plasmodium falciparum*. *J. Struct. Biol.* **177**, 224–232 (2012).
17. Sinden, R. E. The cell biology of sexual development in *Plasmodium*. *Parasitology* **86**, 7–28 (1983).
18. Billker, O. *et al.* Calcium and a Calcium-Dependent Protein Kinase Regulate Gamete Formation and Mosquito Transmission in a Malaria Parasite. *Cell* **117**, 503–514 (2004).
19. Vlachou, D., Schlegelmilch, T., Runn, E., Mendes, A. & Kafatos, F. C. The developmental migration of *Plasmodium* in mosquitoes. *Curr. Opin. Genet. Dev.* **16**, 384–391 (2006).
20. Bozdech, Z. *et al.* The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol.* **1**, 85–100 (2003).
21. Meerstein-kessel, L. *et al.* Probabilistic data integration identifies reliable gametocyte-specific proteins and transcripts in malaria parasites. *Sci. Rep.* 1–13 (2018) doi:10.1038/s41598-017-18840-7.
22. Lasonder, E. *et al.* Integrated transcriptomic and proteomic analyses of *P . falciparum*

gametocytes : molecular insight into sex-specific processes and translational repression. 1–15 (2016) doi:10.1093/nar/gkw536.
23. Callebaut, I., Prat, K., Meurice, E., Mornon, J.-P. & Tomavo, S. Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. *BMC Genomics* **6**, 100 (2005).
24. Adjalley, S. H. *et al.* Landscape and Dynamics of Transcription Initiation in the Malaria Parasite *Plasmodium falciparum* Resource Landscape and Dynamics of Transcription Initiation in the Malaria Parasite Plasmodium falciparum. *CellReports* **14**, 2463–2475 (2016).
25. Chappell, L. *et al.* Refining the transcriptome of the human malaria parasite *Plasmodium falciparum* using amplification-free RNA-seq. *BMC Genomics* **21**, 395 (2020).
26. Watanabe, J., Sasaki, M., Suzuki, Y. & Sugano, S. Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene* **291**, 105–113 (2002).
27. Kensche, P. R. *et al.* The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. *Nucleic Acids Res.* **44**, 2110–2124 (2016).
28. Toenhake, C. G. & Bártfai, R. What functional genomics has taught us about transcriptional regulation in malaria parasites. *Brief. Funct. Genomics* (2019) doi:10.1093/bfgp/elz004.
29. Toenhake, C. G. *et al.* Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying *Plasmodium falciparum* Blood-Stage Development. *Cell Host Microbe* **23**, 557-569.e9 (2018).
30. Ay, F. *et al.* Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.* **24**, 974–988 (2014).
31. Sierra-Miranda, M. *et al.* PfAP2Tel, harbouring a non-canonical DNA-binding AP2 domain, binds to *Plasmodium falciparum* telomeres. *Cell. Microbiol.* **19**, e12742 (2017).
32. Gupta, A. P. & Bozdech, Z. Epigenetic landscapes underlining global patterns of gene expression in the human malaria parasite, *Plasmodium falciparum*. *Int. J. Parasitol.* **47**, 399–407 (2017).
33. Fraschka, S. A. *et al.* Comparative Heterochromatin Profiling Reveals Conserved and Unique Epigenome Signatures Linked to Adaptation and Development of Malaria Parasites. *Cell Host Microbe* **23**, 407-420.e8 (2018).
34. Bischoff, E. & Vaquero, C. In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of *Plasmodium falciparum*. *BMC Genomics* **11**, 34 (2010).
35. Sinha, A. *et al.* A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium* . *Nature* **507**, 253–257 (2014).
36. Tintó-Font, E. *et al.* The PfAP2-HS transcription factor protects malaria parasites from febrile temperatures. *bioRxiv* 2021.03.15.435375 (2021).
37. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**, 248–252 (2014).
38. Josling, G. A. *et al.* Dissecting the role of PfAP2-G in malaria gametocytogenesis. *Nat. Commun.* **11**, 1503 (2020).
39. Yuda, M. *et al.* Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Mol. Microbiol.* **71**, 1402–1414 (2009).
40. Scherf, A. & Lopez-rubio, J. J. Antigenic Variation in *Plasmodium falciparum*. (2008).
41. Singh, S. *et al.* The PfAP2-G2 transcription factor is a critical regulator of gametocyte maturation. *Mol. Microbiol.* (2020).
42. Li, Y., Baptista, R. P. & Kissinger, J. C. Noncoding RNAs in Apicomplexan Parasites:

An Update. *Trends Parasitol.* **36**, 835–849 (2020).
43. Broadbent, K. M. *et al.* Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* **16**, 454 (2015).
44. Poran, A. *et al.* Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* **551**, 95–99 (2017).
45. Filarsky, M. *et al.* GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science (80-.).* **359**, 1259–1263 (2018).
46. Flueck, C. *et al.* A Major Role for the *Plasmodium falciparum* ApiAP2 Protein PfSIP2 in Chromosome End Biology. *PLOS Pathog.* **6**, e1000784 (2010).
47. Gissot, M., Briquet, S., Refour, P., Boschet, C. & Vaquero, C. PfMyb1, a *Plasmodium falciparum* transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. *J. Mol. Biol.* **346**, 29–42 (2005).
48. Painter, H. J. *et al.* Genome-wide real-time in vivo transcriptional dynamics during *Plasmodium falciparum* blood-stage development. *Nat. Commun.* **9**, 1–12 (2018).
49. Painter, H. J., Carrasquilla, M. & Llinás, M. Capturing in vivo RNA transcriptional dynamics from the malaria parasite *Plasmodium falciparum*. *Genome Res.* 1–21 (2017).
50. Shock, J. L., Fischer, K. F. & DeRisi, J. L. Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.* **8**, R134–R134 (2007).
51. Baum, J. *et al.* Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. *Nucleic Acids Res.* **37**, 3788–3798 (2009).
52. Reddy, B. P. N. *et al.* A bioinformatic survey of RNA-binding proteins in *Plasmodium*. *BMC Genomics* 1–26 (2015).
53. Vembar, S. S., Droll, D. & Scherf, A. Translational regulation in blood stages of the malaria parasite *Plasmodium spp*.: systems-wide studies pave the way. *Wiley Interdiscip. Rev. RNA* **7**, 772–792 (2016).
54. Balu, B. *et al.* CCR4-associated factor 1 coordinates the expression of *Plasmodium falciparum* egress and invasion proteins. *Eukaryot. Cell* **10**, 1257–1263 (2011).
55. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
56. Peterson, C. L. & Laniel, M.-A. Histones and histone modifications. *Curr. Biol.* **14**, R546-51 (2004).
57. Bártfai, R. *et al.* H2A.Z Demarcates Intergenic Regions of the *Plasmodium falciparum* Epigenome That Are Dynamically Marked by H3K9ac and H3K4me3. *PLOS Pathog.* **6**, e1001223 (2010).
58. Gupta, A. P. *et al.* Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite *Plasmodium falciparum*. *PLoS Pathog.* **9**, e1003170 (2013).
59. Gupta, A. P. *et al.* Histone 4 lysine 8 acetylation regulates proliferation and host–pathogen interaction in *Plasmodium falciparum*. *Epigenetics Chromatin* **10**, 40 (2017).
60. Salcedo-Amaya, A. M. *et al.* Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9655–60 (2009).
61. Raabe, C. A. *et al.* A global view of the nonprotein-coding transcriptome in *Plasmodium falciparum*. *Nucleic Acids Res.* **38**, 608–617 (2010).
62. Jiang, L. *et al.* PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. *Nature* **499**, 223–227 (2013).
63. Vembar, S. S., Scherf, A. & Siegel, T. N. Noncoding RNAs as emerging regulators of *Plasmodium falciparum* virulence gene expression. *Curr. Opin. Microbiol.* **20**, 153–161 (2014).
64. Amit-Avraham, I. *et al.* Antisense long noncoding RNAs regulate var gene activation in the malaria parasite *Plasmodium falciparum*. *Proc. Natl. Acad. Sci.* **112**, E982 LP-E991 (2015).
65. Davidson, E. & Levin, M. Gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.*

102, 4935 LP – 4935 (2005).
66. Isewon, I., Oyelade, J., Brors, B. & Adebiyi, E. In silico gene regulatory network of the Maurer's cleft pathway in *Plasmodium falciparum*. *Evol. Bioinforma.* **11**, 231–238 (2015).
67. Li, E. & Davidson, E. H. Building developmental gene regulatory networks. *Birth Defects Res. C. Embryo Today* **87**, 123–130 (2009).
68. Singh, A. J., Ramsey, S. A., Filtz, T. M. & Kioussi, C. Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.* **75**, 1013–1025 (2018).
69. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **11**, 559 (2008).
70. Pruitt, K. D. *et al.* An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **29**, 137–140 (2003).
71. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* **5**, e12776 (2010).
72. Villaverde, A. F., Ross, J., Morán, F. & Banga, J. R. MIDER: Network Inference with Mutual Information Distance and Entropy Reduction. *PLoS One* **9**, e96732 (2014).
73. Chai, L. E. *et al.* A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* **48**, 55–65 (2014).
74. Delgado, F. M. & Gómez-Vela, F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artif. Intell. Med.* **95**, 133–145 (2019).
75. Xuan Vinh, N., Chetty, M., Coppel, R. & Wangikar, P. P. Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network. *BMC Bioinformatics* **13**, (2012).
76. Liu, F., Zhang, S.-W., Guo, W.-F., Wei, Z.-G. & Chen, L. Inference of Gene Regulatory Network Based on Local Bayesian Networks. *PLOS Comput. Biol.* **12**, e1005024 (2016).
77. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* **1**, (2007).
78. Moerman, T. *et al.* GRNBoost2 and Arboreto : efficient and scalable inference of gene regulatory networks. *Bioinformatics* **3**, 2–3 (2018).
79. Aibar, S. *et al.* SCENIC : single-cell regulatory network inference and clustering. *Nat. Methods* 1–4 (2017) doi:10.1038/nmeth.4463.
80. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**, 21 (2013).
81. Muzio, G., O'Bray, L. & Borgwardt, K. Biological network analysis with deep learning. *Brief. Bioinform.* **00**, 1–17 (2020).
82. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* (2014).
83. Kc, K., Li, R., Cui, F., Yu, Q. & Haake, A. R. GNE: A deep learning framework for gene network inference by aggregating biological information. *BMC Syst. Biol.* **13**, 1–14 (2019).
84. Saint-Antoine, M. M. & Singh, A. Network inference in systems biology: recent developments, challenges, and applications. *Curr. Opin. Biotechnol.* **63**, 89–98 (2020).
85. Gladitz, J., Klink, B. & Seifert, M. Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion. *Acta Neuropathol. Commun.* **6**, 49 (2018).
86. Xu, T., Ou-Yang, L., Hu, X. & Zhang, X.-F. Identifying Gene Network Rewiring by Integrating Gene Expression and Gene Network Data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **15**, 2079–2085 (2018).
87. Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., Reverter, A. & Ragan, M. A. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med.* **4**, 41 (2012).

88. Vundavilli, H. *et al.* Bayesian Inference Identifies Combination Therapeutic Targets in Breast Cancer. *IEEE Trans. Biomed. Eng.* **66**, 2684–2692 (2019).
89. Allahyar, A., Ubels, J. & de Ridder, J. A data-driven interactome of synergistic genes improves network-based cancer outcome prediction. *PLOS Comput. Biol.* **15**, e1006657 (2019).
90. Hu, G. *et al.* Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nat. Biotechnol.* **28**, 91–98 (2010).
91. Bozdech, Z. *et al.* Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. (2003).
92. Tienda-Luna, I. M. *et al.* Inferring the skeleton cell cycle regulatory network of malaria parasite using comparative genomic and variational Bayesian approaches. *Genetica* **132**, 131–142 (2008).
93. Rono, M. K. *et al.* Adaptation of *Plasmodium falciparum* to its transmission environment. *Nat. Ecol. Evol.* **2**, 377–387 (2018).
94. Van Biljon, R. *et al.* Inducing controlled cell cycle arrest and re-entry during asexual proliferation of *Plasmodium falciparum* malaria parasites. *Sci. Rep.* **8**, 1–14 (2018).
95. van Biljon, R. *et al.* Hierarchical transcriptional control regulates *Plasmodium falciparum* sexual differentiation. *BMC Genomics* **20**, 920 (2019).
96. PlasmoDB. https://plasmodb.org.
97. Otto, T. D. *et al.* New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol. Microbiol.* **76**, 12–24 (2010).
98. Doerig, C. & Tobin, A. B. Previews Parasite Protein Kinases : At Home and Abroad. *Cell Host Microbe* **8**, 305–307 (2010).
99. Doerig, C., Chakrabarti, D., Kappes, B. & Matthews, K. The cell cycle in protozoan parasites. *Prog. Cell Cycle Res.* **4**, 163–183 (2000).
100. Doerig, C., Endicott, J. & Chakrabarti, D. Cyclin-dependent kinase homologues of *Plasmodium falciparum*. *Int. J. Parasitol.* **32**, 1575–1585 (2002).
101. Francia, M. E. & Striepen, B. Cell division in apicomplexan parasites. *Nat. Rev. Microbiol.* **12**, 125–136 (2014).
102. Striepen, B., Jordan, C. N., Reiff, S. & van Dooren, G. G. Building the Perfect Parasite: Cell Division in Apicomplexa. *PLOS Pathog.* **3**, e78 (2007).
103. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507**, 248–52 (2014).
104. Llorà-Batlle, O. *et al.* Conditional expression of PfAP2-G for controlled massive sexual conversion in *Plasmodium falciparum*. *Sci. Adv.* **6**, eaaz5057–eaaz5057 (2020).
105. Young, J. A. *et al.* The *Plasmodium falciparum* sexual development transcriptome : A microarray analysis using ontology-based pattern identification. **143**, 67–79 (2005).
106. Campbell, T. L., de Silva, E. K., Olszewski, K. L., Elemento, O. & Llinás, M. Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6**, (2010).
107. Balaji, S., Babu, M. M., Iyer, L. M. & Aravind, L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.* **33**, 3994–4006 (2005).
108. Painter, H. J., Campbell, T. L. & Llinas, M. The Apicomplexan AP2 family: Integral factors regulating Plasmodium development. *Mol. Biochem. Parasitol.* **176**, 1–7 (2011).
109. Morrissey, E. R. GRENITS : Gene Regulatory Network Inference Using Time Series Example : Network Inference For Simulated Data. 1–5 (2011).
110. Trager, W. & Jensen, J. B. Human Malaria Parasites in Continuous Culture. *Science (80-.).* **193**, 673–675 (1976).
111. Allen, R. J. W. & Kirk, K. *Plasmodium falciparum* culture : The benefits of shaking. *Mol. Biochem. Parasitol.* **169**, 63–65 (2010).
112. van Biljon, R. Integrative transcriptome and phenome analysis reveal unique regulatory cascades controlling the intraerythrocytic asexual and sexual development of human malaria parasites. (University of Pretoria, 2017).

113. Szklarczyk, D. *et al.* STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
114. Adjalley, S. H. *et al.* Quantitative assessment of *Plasmodium falciparum* sexual development reveals potent transmission- blocking activity by methylene blue. *PNAS* **108**, E1214–E1223 (2011).
115. Reader, J. *et al.* Nowhere to hide: interrogating different metabolic parameters of *Plasmodium falciparum* gametocytes in a transmission blocking drug discovery pipeline towards malaria elimination. *Malar. J.* **14**, 1–17 (2015).
116. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
117. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
118. Pandey, R. *et al.* Genome wide in silico analysis of *Plasmodium falciparum* phosphatome. *BMC Genomics* **15**, 1024 (2014).
119. Cui, L. & Miao, J. Chromatin-Mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. *Eukaryot. Cell* **9**, 1138–1149 (2010).
120. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).
121. Matesanz, F., Téllez, M. D. M. & Alcina, A. The *Plasmodium falciparum* fatty acyl-CoA synthetase family (PfACS) and differential stage-specific expression in infected erythrocytes. *Mol. Biochem. Parasitol.* **126**, 109–112 (2003).
122. Bethke, L. L. *et al.* Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **150**, 10–24 (2006).
123. Lamour, S. D., Straschil, U., Saric, J. & Delves, M. J. Changes in metabolic phenotypes of *Plasmodium falciparum* in vitro cultures during gametocyte development. *Malar. J.* **13**, 1–10 (2014).
124. Hliscs, M. *et al.* Organization and function of an actin cytoskeleton in *Plasmodium falciparum* gametocytes. *Cell. Microbiol.* **17**, 207–225 (2015).
125. Mair, G. R. *et al.* Universal features of post-transcriptional gene regulation are critical for *Plasmodium* zygote development. *PLoS Pathog.* **6**, (2010).
126. Miao, J. *et al.* Puf Mediates Translation Repression of Transmission-Blocking Vaccine Candidates in Malaria Parasites. *PLoS Pathog.* **9**, (2013).
127. Miao, J. *et al.* The Puf-family RNA-binding protein PfPuf2 regulates sexual development and sex differentiation in the malaria parasite *Plasmodium falciparum*. *J. Cell Sci.* **123**, 1039–1049 (2010).
128. Daire, V. & Poüs, C. Kinesins and protein kinases: Key players in the regulation of microtubule dynamics and organization. *Arch. Biochem. Biophys.* **510**, 83–92 (2011).
129. Ginsburg, H. & Tilley, L. *Plasmodium falciparum* metabolic pathways (MPMP) project upgraded with a database of subcellular locations of gene products. *Trends Parasitol.* **27**, 285–286 (2011).
130. Carvalho, T. G. *et al.* The ins and outs of phosphosignalling in *Plasmodium*: Parasite regulation and host cell manipulation. *Mol. Biochem. Parasitol.* **208**, 2–15 (2016).
131. Reininger, L. *et al.* A NIMA-related protein kinase is essential for completion of the sexual cycle of malaria parasites. *J. Biol. Chem.* **280**, 31957–31964 (2005).
132. Dearnley, M. K. *et al.* Origin, composition, organization and function of the inner membrane complex of *Plasmodium falciparum* gametocytes. *J. Cell Sci.* **125**, 2053–2063 (2012).
133. Sinden, R. E. & Smalley, M. E. Gametocytogenesis of *Plasmodium falciparum* in vitro: the cell-cycle. *Parasitology* **79**, 277–296 (1979).
134. Modrzynska, K. *et al.* A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the *Plasmodium* Life Cycle. *Cell Host Microbe* **21**, 11–22 (2017).
135. Ginsburg, H. Progress in in silico functional genomics : the malaria Metabolic

Pathways database. *Trends Parasitol.* **22**, 238–240 (2006).
136. Aikawa, M., Huff, C. G. & Sprinz, H. Comparative Fine Structure Study of the Gametocytes of Avian , Reptilian , ond Mammalian Malarial Parasites. *J. Ultrastruct. Res.* **331**, 316–331 (1969).
137. Bounkeua, V., Li, F. & Vinetz, J. M. In Vitro Generation of *Plasmodium falciparum* Ookinetes. **83**, 1187–1194 (2010).
138. Lobo, C. a & Kumar, N. Sexual differentiation and development in the malaria parasite. *Parasitol. Today* **14**, 146–50 (1998).
139. Kumar, N. & Carter, R. Biosynthesis of the target antigens of antibodies blocking transmission of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **13**, 333–342 (1984).
140. Vermeulen, A. N. *et al.* Characterization of *Plasmodium falciparum* sexual stage antigens and their biosynthesis in synchronised gametocyte cultures. *Mol. Biochem. Parasitol.* **20**, 155–163 (1986).
141. Rittershaus, E. S. C., Baek, S. H. & Sassetti, C. M. The normalcy of dormancy: Common themes in microbial quiescence. *Cell Host Microbe* **13**, 643–651 (2013).
142. Quan, Z. *et al.* The Yeast GSK-3 Homologue Mck1 Is a Key Controller of Quiescence Entry and Chronological Lifespan. *PLoS Genet.* **11**, e1005282 (2015).
143. Romer, T. H. *et al.* The Structure and Performance of Interpreters. in *Proceedings of the Seventh International Conference on Architectural Support for Programming Languages and Operating Systems* 150–159 (Association for Computing Machinery, 1996).
144. Josling, G. & Llinás, M. Sexual development in *Plasmodium* parasites: knowing when it's time to commit. *Nat. Rev. Microbiol.* **13**, 573–587 (2015).
145. Abel, S. & Le Roch, K. G. The role of epigenetics and chromatin structure in transcriptional regulation in malaria parasites. *Brief. Funct. Genomics* (2019).
146. Sardar, R. *et al.* ApicoTFdb: The comprehensive web repository of apicomplexan transcription factors and regulators. *bioRxiv* 530006 (2019).
147. Coetzee, N. *et al.* Quantitative chromatin proteomics reveals a dynamic histone post-translational modification landscape that defines asexual and sexual *Plasmodium* falciparum parasites. *Sci. Rep.* **7**, 607 (2017).
148. Reid, A. J. *et al.* Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7**, e33105 (2018).
149. López-barragán, M. J. *et al.* Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. (2011).
150. LaCount, D. J. *et al.* A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **438**, 103–107 (2005).
151. Sorber, K., Dimon, M. T. & Derisi, J. L. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions , alternative splicing and splicing of antisense transcripts. **39**, 3820–3835 (2011).
152. Hoeijmakers, W. A. M. *et al.* H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the *Plasmodium falciparum* genome. *Mol. Microbiol.* **87**, 1061–1073 (2013).
153. Siegel, T. N. *et al.* Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. (2014).
154. Carrington, E. *et al.* The ApiAP2 factor PfAP2-HC is an integral component of heterochromatin in the malaria parasite *Plasmodium falciparum bioRxiv* 2020.11.06.370338 (2020).
155. Broadbent, K. M. *et al.* A global transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs. *Genome Biol.* **12**, R56 (2011).
156. Adjalley, S. H. *et al.* Quantitative assessment of Plasmodium falciparum sexual development reveals potent transmission-blocking activity by methylene blue. *Proc. Natl. Acad. Sci.* **108**, E1214–E1223 (2011).
157. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with

RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
158. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 4–8 (2016).
159. Love MI, Huber W, A. S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).
160. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, S. G. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43(7)**, (2015).
161. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Inter J Complex Syst* **1695**, (2006).
162. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).
163. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
164. Dai, R., Xia, Y., Liu, C. & Chen, C. csuWGCNA: a combination of signed and unsigned WGCNA to capture negative correlations. *bioRxiv* 288225 (2019).
165. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).
166. Tilley, L., Dixon, M. W. A. & Kirk, K. The *Plasmodium falciparum*-infected red blood cell. *Int. J. Biochem. Cell Biol.* **43**, 839–842 (2011).
167. Silvestrini, F. *et al.* Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum.* **143**, 100–110 (2005).
168. Li, Z. *et al. Plasmodium* transcription repressor AP2-O3 regulates sex-specific identity of gene expression in female gametocytes. *EMBO Rep.* **22**, e51660 (2021).
169. Siciliano, G. *et al.* A high susceptibility to redox imbalance of the transmissible stages of *Plasmodium falciparum* revealed with a luciferase-based mature gametocyte assay. *Mol. Microbiol.* **104**, 306–318 (2017).
170. Lee, M. C. S., Lindner, S. E., Lopez-Rubio, J.-J. & Llinás, M. Cutting back malaria: CRISPR/Cas9 genome editing of *Plasmodium*. *Brief. Funct. Genomics* **18**, 281–289 (2019).
171. Santos, J. M. *et al.* Red Blood Cell Invasion by the Malaria Parasite Is Coordinated by the PfAP2-I Transcription Factor. *Cell Host Microbe* **21**, 731-741.e10 (2017).
172. Wiegant, J. C. A. G., Dirks, R. W., Dimopoulos, G., Janse, C. J. & Andrew, P. Universal Features of Post-Transcriptional Gene Regulation Are Critical for *Plasmodium* Zygote Development. **6**, (2010).
173. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
174. Elemento, O., Slonim, N. & Tavazoie, S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Mol. Cell* **28**, 337–350 (2007).
175. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
176. Tao, D. *et al.* A saliva-based rapid test to quantify the infectious subclinical malaria parasite reservoir. *Sci. Transl. Med.* **11**, eaan4479 (2019).
177. Shrestha, S., Li, X., Ning, G., Miao, J. & Cui, L. The RNA-binding protein Puf1 functions in the maintenance of gametocytes in *Plasmodium falciparum*. *J. Cell Sci.* **129**, 3144 LP – 3152 (2016).
178. Nixon, C. P. *et al.* Antibodies to PfsEGXP, an Early Gametocyte-Enriched Phosphoprotein, Predict Decreased Plasmodium falciparum Gametocyte Density in Humans. *J. Infect. Dis.* **218**, 1792–1801 (2018).
179. Zhang, M. *et al.* Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. *Science (80-.).* **506**, (2018).
180. Shannon, P. *et al.* Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
181. Sinha, A. *et al.* A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. *Nature* **507**, 253–7 (2014).

182. Stubbs, J. *et al.* Molecular mechanism for switching of *P. falciparum* invasion pathways into human erythrocytes. *Science* **309**, 1384–1387 (2005).
183. Yu, F., Yang, S., Li, Y. & Hu, W. Co-expression network with protein – protein interaction and transcription regulation in malaria parasite *Plasmodium falciparum*. **518**, 7–16 (2013).
184. Zhang, B., Horvath, S., Zhang, B. & Horvath, S. Statistical Applications in Genetics and Molecular Biology A General Framework for Weighted Gene Co-Expression Network Analysis A General Framework for Weighted Gene Co-Expression Network Analysis ∗. **4**, (2005).
185. Yılmaz Isıkhan, S., Karabulut, E. & Alpar, C. R. Determining Cutoff Point of Ensemble Trees Based on Sample Size in Predicting Clinical Dose with DNA Microarray Data. *Comput. Math. Methods Med.* **2016**, 6794916 (2016).
186. JetBrains. Python Developers Survey 2018 Results. *www.jetbrains.com* https://www.jetbrains.com/research/python-developers-survey-2018/ (2018).
187. Hollin, T. & Le Roch, K. G. From Genes to Transcripts, a Tightly Regulated Journey in Plasmodium. *Front. Cell. Infect. Microbiol.* **10**, 801 (2020).
188. Bunnik, E. M. *et al.* Comparative 3D genome organization in apicomplexan parasites. *Proc. Natl. Acad. Sci.* **116**, 3183 LP – 3192 (2019).
189. Amit-Avraham, I. *et al. Plasmodium falciparum* erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to CD36, thrombospondin, and intercellular adhesion molecule 1. *PNAS* **93**, 3497–3502 (2015).